36-402 DA Exam 2

Keltin Grimes (kgrimes)

May 6, 2022

Introduction

Providing loans to small businesses is a risky endeavor for banks due to the high rate of failure in small or newly formed businesses. There is a large potential upside for the economy, however, as successful small business can grow rapidly, adding many new jobs as they do so. The United States federal government, therefore, has a strong interest in promoting the success of small businesses in the US. Our client, the US Small Business Administration (SBA), is a federal agency designed to support small businesses, create jobs, and boost the economy. The SBA has a loan program in which the federal government will guarantee loans to small businesses, so if a business defaults on a loan given by a bank, the bank is not liable for the entire loan amount. This lessens the risk of the bank and makes them more likely to offer loans to small businesses. The SBA would like to optimize their expenditures to create as many jobs given limited funding. **(1)** Therefore in this report we have aimed to analyze the relationship between jobs created and dollars loaned, and investigate which types of businesses create the most jobs per dollars loaned.

(2) We examined two models of jobs created per dollars loaned using four metrics of business type: one where the relationship between jobs and loan amount was constrained to be linear, and one where the relationship was allowed to be non-linear. We found that the non-linear model was significantly better, and that generally the number of jobs created increased with dollars loaned, up until loan amounts around $1 million, when jobs created actually decreased. We also found that businesses in the industries of `AccommodationFood`, `AdminSupport`, and `MiningGas` had the highest job creation rates, controlling for the other variables. We therefore recommend the SBA to not guarantee loans much over $1 million, and to focus on business in the industries previously mentioned, but admit that our analysis comes with significant limitations, which we discuss.

Exploratory Data Analysis

We have a dataset of information on 16,245 SBA loans granted to businesses in Pennsylvania between 1995 and 2014. The data consists of 20 variables, but we only use a subset of these for our analysis. After removing observations with missing values, we end up with 15855 unique loans. Since we are interested in comparing job-creation rates for different types of businesses, we examine the following four variables: whether the business is urban or rural, whether or not it is a new business, whether or not it is a franchise, and the industry of the business. There are 19 different industry categories, including `Other`. **(1)** There are 13,461 loans to urban businesses and 2,394 to rural businesses; there are 3,738 loans to new businesses and 12,117 to existing ones; and there are 459 loans to franchises and 15,396 to non-franchises. The distribution of industries is shown in Figure 1. The largest industry by a good amount is `RetailTrade`, but the `ProfServices`, `Other`, and `Construction` industries are also quite large. The smallest industries are `PublicAdmin`, `Utilities`, and `MiningGas`, in that order, and all three have less than 25 observations. We also need to be careful to include any confounding variables we have access to. Based on our prior knowledge we would expect larger loans to create more jobs because with more money a business would be able to hire more people. We also want to create the most jobs while spending the least money, so specifically what we care about is the rate of jobs created per dollar loaned. We will model this by including the loan amount as a covariate in any models we create. In Figure 2 we show the marginal distribution of loan amount (left). We find that it is unimodal and centered at a mean of $130,607, but is extremely right skewed.

(2) Our variable of interest is the number of jobs the applicant business expects to create using the loan they are applying for. We plot the marginal distribution of this estimate of jobs created in Figure 2 (right). Similar to the distribution of loan amount, it is unimodal with extreme right skew, and centered at a mean of 1.383 jobs. There are 11,388 observations that report 0 jobs created, and another 2,420 that report either 1 or 2 jobs. The large amount of zero observations suggests that modeling will be difficult. There is also a noticeable outlier with 320 expected jobs created, nearly double the amount of the next largest observation, so we will be sure to check for outliers when creating out models. **(3)** After exploring the relationship between jobs created and loan amount and looking at various transformations of the two variables, we found the most linear relationship to be when applying a log transform to both values. Applying a log transformation to jobs created is useful because it corresponds nicely with Poisson regression, as we discuss in the next section. We plot this relationship in Figure 3. Note that we add 1 to each observation for

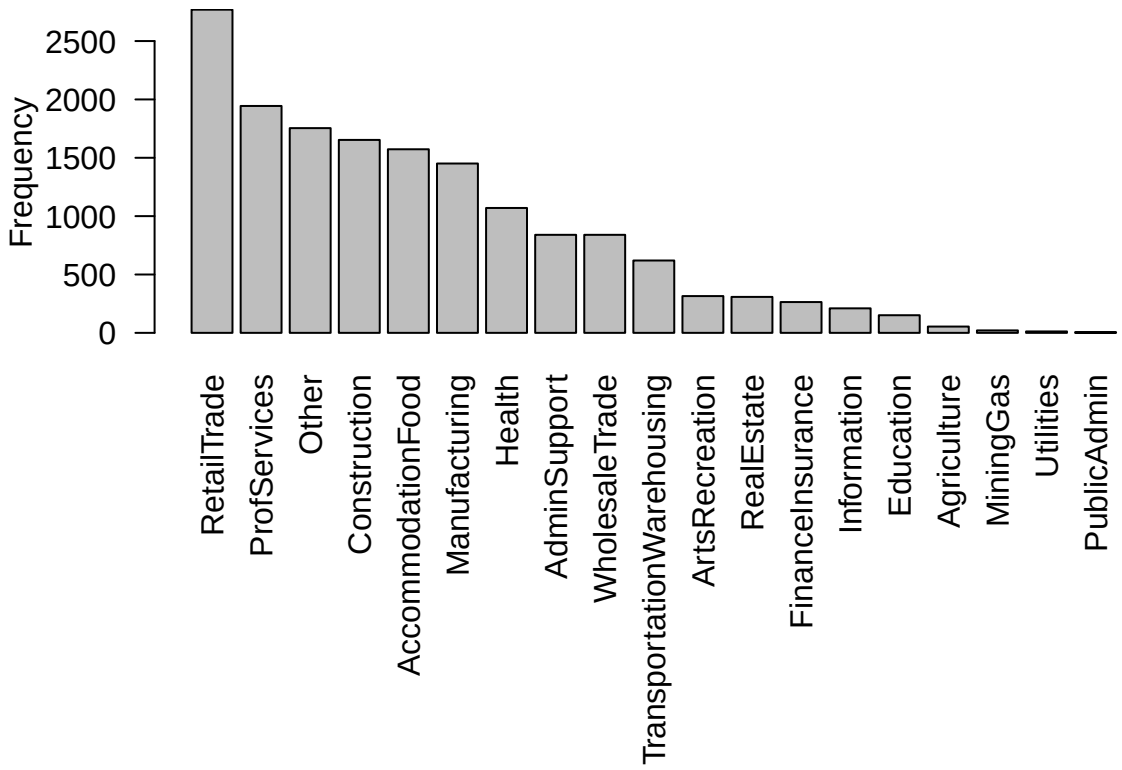

Figure 1: Barplot of Loans Made in Various Business Industries

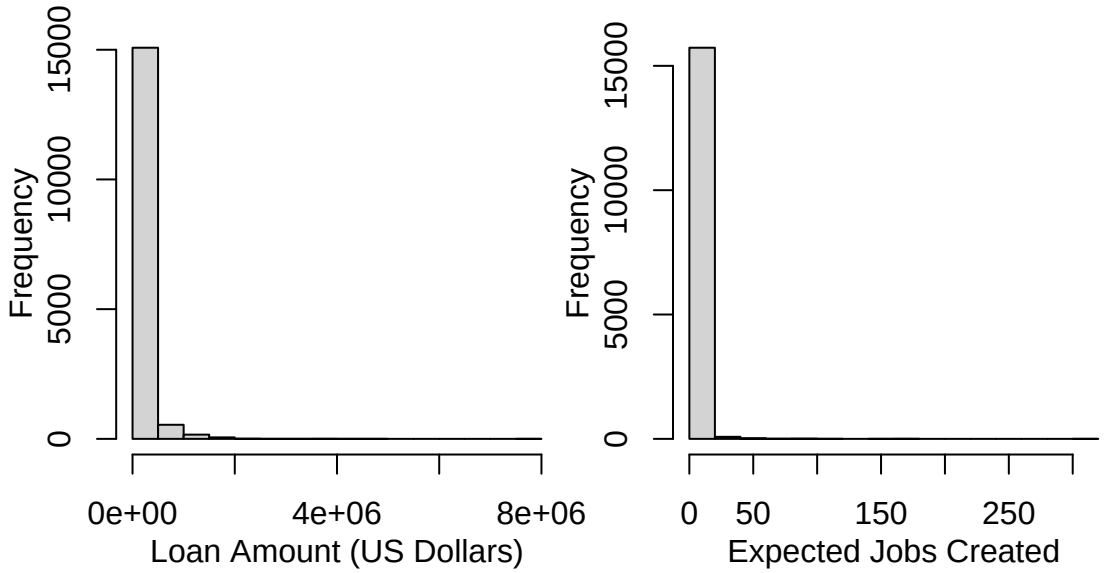

Figure 2: Histograms of Loan Amount in US Dollars (left) and Expected Jobs Created (right).

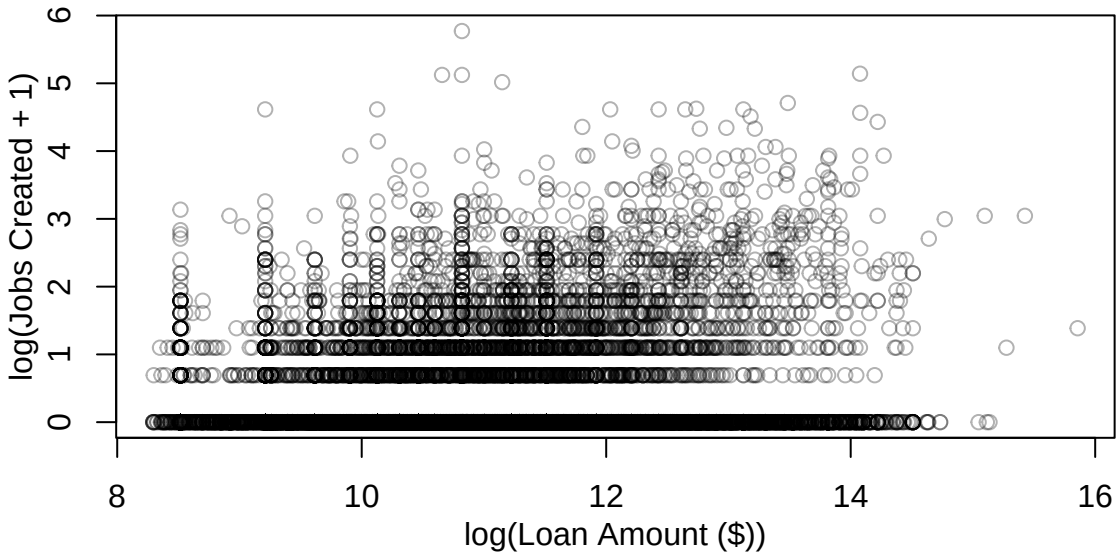

Figure 3: Scatterplot of log(Loan Amount ($)) vs. log(Jobs Created + 1)

jobs created so that we can apply the log transform on the zero values. The points are semi-transparent due to the amount of overlap. We can see that jobs created tends to increase with loan amount, but the relationship is not that strong, and is somewhat distorted by the many points with zero jobs created. It is also not obvious that the relationship is in fact linear. **(4)** The marginal distributions of jobs created given the business-related covariates are difficult to assess due to the number of observations with zero jobs created, but it appears that generally new business create more jobs than existing ones, franchises create more jobs than non-franchises, and urban and rural businesses are roughly the same. The industries that create the most jobs appear to be `AccommodationFood`, `MiningGas`, and `Utilities`.

Modeling & Diagnostics

Our goal is to model the number of jobs created by SBA loans, but jobs are reported as count data, so are not ideally suited for direct linear modeling. **(1)** Instead of using linear regression, we use Poisson regression, or a Generalized Linear Model where we assume that our response distribution is Poisson distributed. We use a logarithmic link function, so our model predicts the log of created jobs. We are interested in the number of jobs created per dollar loaned, and we found in the EDA section that the relationship between log jobs and log loan amount is the most linear, so we include the log-transformed dollar amount of the loan as a covariate. We would like to understand how job creation rates

differ across various types of businesses, so we also include covariates that illustrate the type of business applying for the loan, namely: whether the business is urban or rural, whether or not it is a new business, whether or not it is a franchise, and the industry of the business. **(2)** Since in our EDA we were unable to definitively say whether the relationship between jobs created and loan amount is linear, we will also fit a Generalized Additive Model where the partial response function for log loan amount is a smoothing spline with smoothing parameter chosen by generalized cross-validation with a maximum EDF of 5, keeping everything else the same. Comparing these two models will give us information on whether or not the relationship is in fact linear.

After fitting the two models, we found there to be some very large outliers. We iteratively fitted the models and removed the point with the largest Cooks Distance until all points had a Cooks Distance less than 1. We removed three points in this process. In Figure 4 we show the plot of Cooks Distances for the GLM model before and after removing the outliers. The plots look very similar for the GAM model. Generally we would remove any point with Cooks Distance close to 1, but after removing these three points there was group of 6 points with distance around 0.7, and we determined it best to keep these points rather then potentially removing an important source of information. In Figure 5 we plot the residuals for our two models. The points are partially transparent so the distribution of points is easier to see. **(3)** The residuals appear to roughly be centered at zero for both the GLM and GAM, although it is less obvious at the extremes of loan amount and fitted values. The distribution of the residuals is clearly not constant, however, which calls into question our assumptions about the distribution of our data. Deviations from our assumed model will impact the validity of tests relying on those assumption, so we will have to take that into consideration.

(4) To test whether the non-linear term in the GAM is necessary we use a deviance test, where the test statistic is the difference in deviance between the two models. Since the fitting process for our GAM involves choosing a smoothing parameter with generalized cross-validation, we will use bootstrap to conduct this test. Since our null hypothesis is that the GLM is the correct model, we use parametric bootstrap samples that sample the response values from a Poisson distribution where the mean is the fitted value from the GLM model. Our alternative hypothesis is that the GAM is the correct model. We fit both models (including the GCV process) on 100 bootstrap samples, and calculate the difference in deviance between the two models for each sample. **(5)** We would also like to construct confidence intervals for the coefficients of our models. Again, we must use bootstrap because of our use of GCV to select the smoothing parameter in the GAM. Furthermore,

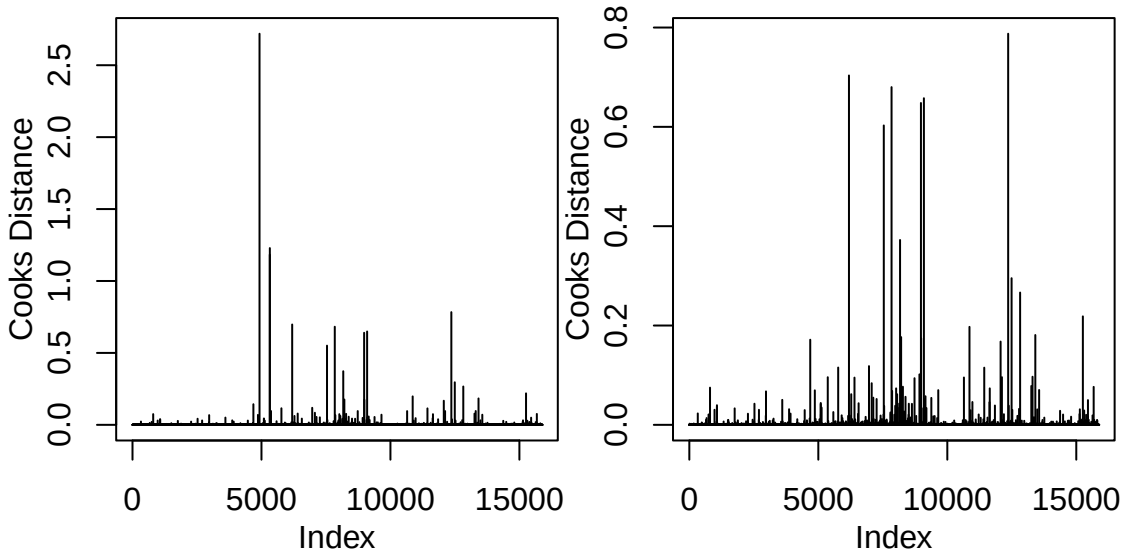

Figure 4: Cooks Distance from GLM before (left) and after (right) removing outliers.

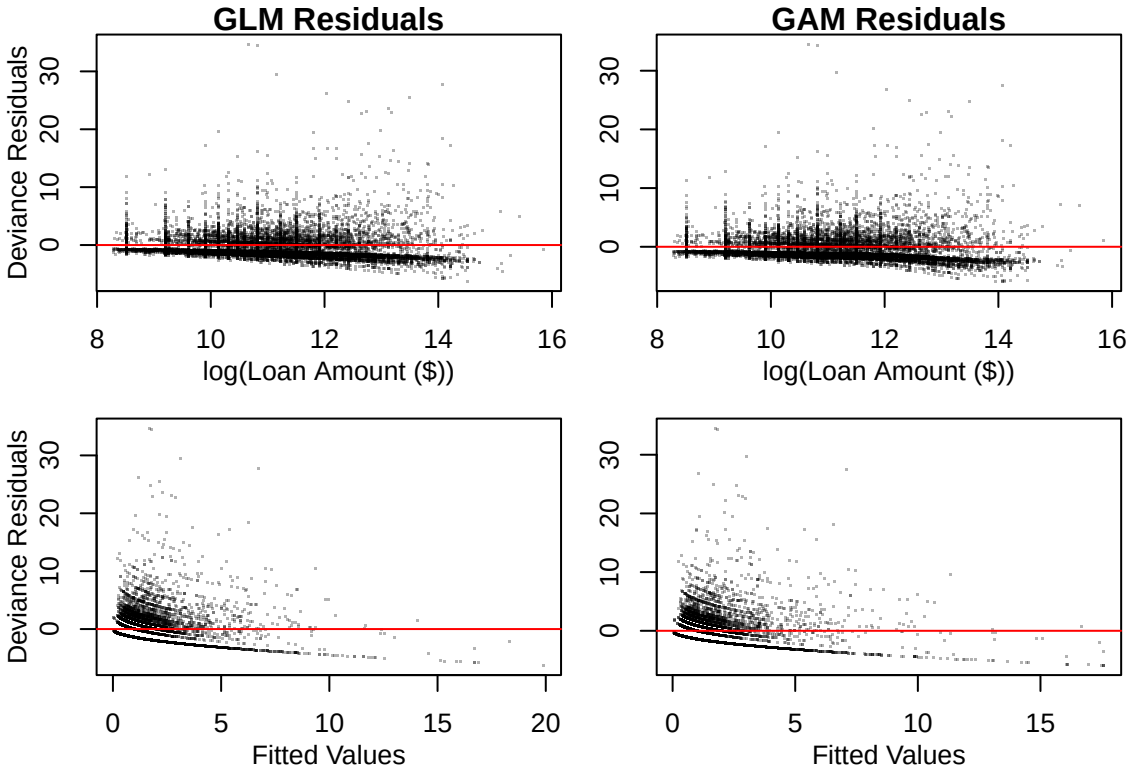

Figure 5: Residuals for GLM (right column) and GAM (left column). Deviance residuals are plotted versus log Loan Amount (top row) and the fitted values (bottom row).

since we are not very confident in the fit of our model, specifically the errors do not seem to be I.I.D., we will bootstrap by resampling cases. We will construct 100 bootstrap samples, fit our models on each sample, record the coefficients for our covariates, and create pivotal confidence intervals for each coefficient.

Results

(1) Our deviance test between the GLM and GAM found a p-value of approximately zero, which means we have significant evidence to reject the null hypothesis that the GLM is the correct model, in favor of the alternative hypothesis that the GAM is. Therefore we will use the GAM for the rest of our analysis. In Figure 6, we have plotted the partial response function for the log of loan amount, and we can see that the function is very non-linear, which supports the results of our deviance test. From this it appears that jobs created increases as loan amount increases, up until loan amounts around 1 million dollars ($\sim e^{14}$), where jobs created starts to decrease. Remember that the y-axis of this plot is in the link scale.

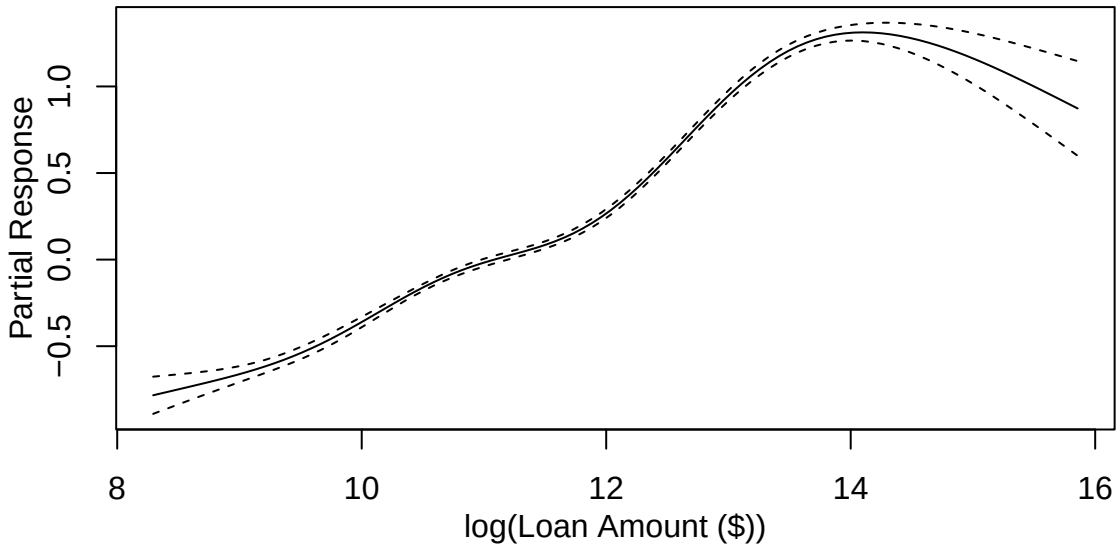

Figure 6: Partial response function for Loan Amount in the GAM.

Table 1: 95% Pivotal Confidence intervals for the coefficient of the industry categories.

	2.5 %	97.5 %
AdminSupport	-0.4578034	-0.0608384
Agriculture	-3.2018741	-2.1447151
ArtsRecreation	-0.7543144	-0.3840657
Construction	-0.9198306	-0.7104745
Education	-1.1122061	-0.3784660
FinanceInsurance	-1.3216315	-0.8247138
Health	-0.8250379	-0.5648205
Information	-1.2887369	-0.8265072
Manufacturing	-0.8429153	-0.5038470
MiningGas	-0.9672815	1.0445580
Other	-1.1593470	-0.8242490
ProfServices	-0.8245396	-0.3124867
PublicAdmin	-3.8276166	18.7112493
RealEstate	-1.7297408	-1.0425166
RetailTrade	-0.9878479	-0.8680020
TransportationWarehousing	-1.0285410	-0.4263729
Utilities	-1.7153217	-0.3920688
WholesaleTrade	-1.3099159	-0.8988123

(2) In Table 1 we present the 95% confidence intervals for the coefficients of the businesses industry. **(3)** The base category is `AccommodationFood`, so for each coefficient β_i we interpret it as: if the business is in category i, we expect that business to create e^{β_i} times as many jobs as a business in `AccommodationFood`, controlling for the other covariates. We can see that none of the confidence intervals are entirely above zero, so no industries create significantly more jobs than `AccommodationFood`. `AdminSupport`, `MiningGas`, and `PublicAdmin` are the only industries not significantly worse than `AccommodationFood`, although the uncertainty for `PublicAdmin` is extremely high. `Agriculture` seems to have the lowest rate of job creation, and `RealEstate`, `FinanceInsurance`, and `WholesaleTrade` also appear to have fairly low rates.

Conclusions

(1) Our report found that, as we would expect, larger loans create more jobs, but we found that there are diminishing returns with loans larger than around $1 million. We therefore recommend the SBA to avoid guaranteeing loans larger than $1 million. We also found, controlling for whether a business is urban or rural, new or existing, and a franchise or not, that businesses in the `AccommodationFood`, `AdminSupport`, and `MiningGas` industries tend to have the highest jobs created per dollar loaned rates. The SBA should prioritize loans to businesses in these industries, and also avoid loans in the `RealEstate`, `FinanceInsurance`, and `WholesaleTrade` industries because of their low job per dollar rates. **(2)** We note that our analysis is not casual, so we cannot say for certain that taking these actions will necessarily result in a more efficient use of the SBA's funding. There are likely many confounding variables that will impact the results of any changes in SBA policy. Additionally, the way in which jobs created is measured in the study is flawed. We are interested in the number of jobs actually created, but only have access to the number of jobs expected to be created as determined by the applicant. These two quantities do not necessarily align, and is open to abuse where an applicant could maliciously overestimate the number of jobs they expect to create to increase their chances of receiving a loan. The SBA may also be interested in the number of jobs retained, as losing jobs can be detrimental to the economy, but this was not analyzed in this report. Another important factor is whether the business defaulted on their loan or payed it back in full. It is much better for the SBA to give loans to businesses that are less likely to default on their loan, even if those businesses tend to create fewer new jobs, because the SBA loses money any time one of the loans they guarantee is defaulted on. Future work could model the probability of a business defaulting, and then we could include a prediction of the likelihood of default as a covariate in the sort of models we examined in this report.