How Effective was the Paycheck Protection Program?

Abhishek Devarajan

November 30, 2022

1 Introduction

The COVID-19 Pandemic hit the United States in March 2020. The rising number of cases, as well as the various lockdown measures implemented at state and federal levels lead to a massive rise in unemployment across the country. 3.3 million Americans filed for unemployment insurance in the week of March 21,2020, a number that increased to 26 million within the following five weeks [1].

In an attempt to mitigate the rise in unemployment and to prevent firm closures, the US federal government, via the Small Business Administration (SBA), launched the Paycheck Protection Program (PPP) in April 2020. The goal of the program was to provide timely relief to small businesses in the form of PPP loans. These loans were meant to help businesses keep up with employee payroll costs, and were forgiven by the SBA given that they went towards employee payrolls.

This paper aims to understand how effective the PPP was at curbing the rise of unemployment. Now that nearly 2 years have past since the onset of COVID in the US, there are ample datasets available to analyze whether PPP lending actually did slow the growth of unemployment. Understanding the strengths and weaknesses of this program can inform policymakers in the US and other countries on how to respond to future economic disruptions.

In Section 2, we will go over the various datasets used to perform our analysis as well as the key variables used in our models. In Section 3, we will go over the techniques used to analyze the data. In Section 4, we will state the results of our models. Finally, in Section 5, we will interpret the findings and suggest future steps for this research.

2 Data

2.1 Sources

The data used in this project is shown in Table 1.

The data was collected from the websites of these respective organizations and processed via Python. The scripts used to collect and clean the data are shown in Section 7.

2.2 Variable Definitions

2.2.1 Change in Unemployment (2020)

Given that this project is estimating the effect of PPP lending on unemployment, we define a variable $\Delta U_{k,j}$ that measures the change in unemployment within year k in a given county, j. More formally, we define

$$\Delta U_{k,j} = \frac{\sum_{i=2}^{4} U_{(k,i),j}}{3} - U_{(k,1),j}, \qquad (1)$$

where $U_{(k,i),j}$ is the unemployment rate in quarter i of year k in county j. In other words, we are measuring how the average unemployment rate across quarters 2 through 4 differs from the unemployment rate in quarter 1. For this project, we will use $\Delta U_{2020,j}$ as our response. The justification for this response comes from the fact that quarter 1 of 2020 occurred before the pandemic and, thus, serves as a control for unemployment levels. We average quarters 2 through 4 in order to get a single number that illustrates COVID-level unemployment. Taking the difference between these values roughly indicates the effect of the pandemic on a county's unemployment rate within a given year.

Using kernel density estimation(KDE) (see Section 3), we can illustrate the distribution of $\Delta U_{k,j}$ across all counties for various years. This illustration can be seen in Figure 1.

In the graph above, we can see that the pre-pandemic distribution of change in unemployment, given by $\Delta U_{2019,j}$ has very little spread and a slight negative skew. Intuitively, this negative skew is due to seasonal jobs such as farming or construction. Typically, these jobs occur in later quarters but not in quarter 1. Looking at the green line in Figure 1, we can see that the distribution of $\Delta U_{2021,j}$ is very close to the distribution of $\Delta U_{2019,j}$. This seems to indicate that the labor market successfully rebounded to pre-pandemic levels in 2021. The clear outlier in the graph is $\Delta U_{2020,j}$, which has a much higher spread

Data	Source
PPP Loan Data	Small Business Administration
Monthly Unemployment Rate by County	St. Louis Federal Reserve
COVID Cases/Deaths by County	Centers for Disease Control & Prevention

Table 1: Datasets and Sources

and is centered around a positive value, rather than a negative one. We posit that this drastic change in distribution over the course of one year is mostly due to the economic disruptions caused by COVID-19.

2.2.2 Forgiven Loan Amount

Next, we will construct a small set of predictors to regress against our response. Most crucially, we need to include a variable that measures the amount of PPP lending received by a given county. For this purpose, we have two obvious choices:

1. T_j: The total amount of PPP loan money received by county j ($ Billions)

2. F_j: The amount of PPP loan money in county j forgiven by the US government ($ Billions)

Interestingly, in Figure 2, we can see that the estimated distribution of T_j and F_j are nearly identical.

This seems to indicate that almost every PPP loan issued was eventually forgiven, regardless of factors such as firm size, loan amount, or industry type. Given this information, we conclude that we can choose to include T_j or F_j in our model without any significantly different results. We opt for including F_j due to the criteria for forgiveness outlined by the SBA. Specifically, since PPP loan forgiveness indicates that businesses used the loan money to pay their employees, we would expect the forgiven lending amount to have a more direct impact on change in unemployment.

2.2.3 Rural Lending Percentage

From here, we examine the distribution of PPP loans across the United States.

Figure 3 shows the spatial distribution of PPP loans across the country. Note that the distribution is very similar to the urban population density of the United States, shown in Figure 4.

Since the PPP loans were mostly distributed in urban areas, we cannot be sure whether the effect of F_j on $\Delta U_{2020,j}$ is due to the forgiven loan amount, or due to some other characteristic of urban/rural counties. To avoid this confusion, we introduce the variable R_j which indicates the rural lending percentage of county j. More specifically, R_j denotes the percentage of PPP loans in county j that were given to businesses that the SBA denoted as "rural".

2.2.4 Change in Unemployment (2019)

Following Formula (1), we construct the variable $\Delta U_{2019,j}$ and use it as a predictor. The rationale behind including this variable as a predictor is to account for any cyclical unemployment patterns that occur consistently between years.

2.2.5 Total Covid Cases/Deaths

Finally, we construct the variables C_j and D_j which represent the total COVID cases and the total COVID deaths in county j. Note that we cannot include both of these variables in the same model, due to their high correlation with one another. Instead, we will fit multiple models, swapping between C_j and D_j. These variables serve as a way to control for the differing levels of COVID infections across different counties.

3 Methods

3.1 Kernel Density Estimation

The first major technique we used in our analysis was Kernel Density Estimation (KDE). Kernel Density Estimation is a non-parametric method that is used to approximate the underlying distribution of data given a sample. In this project, we utilize KDE in both 1-dimensional and 2-dimensional contexts, in order to create various figures.

In the 1D case, KDE can be represented using the following equation

$$\hat{f}(z) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{z - x_i}{h}\right) \tag{2}$$

where $(x_1, \ldots, x_n)$ are i.i.d. samples from an unknown distribution. The variable h represents the "bandwith", and is used to control the convergence and smoothness of the KDE. Finally $\hat{h}$ represents the estimated PDF of the distribution of $(x_1, \ldots, x_n)$.

For our work, we implement Equation (2) in C (see Section 7.3).

For the 2D case, we utilize the function `kde2d` from the R package "MASS". This technique was used to generate Figure 3.

3.2 Kernel Regression

Kernel Regression is a type of non-parametric regression that is used to estimate the relationship between two variables. Specifically, for two random variables X and Y with samples $(x_1, \ldots, x_n)$ and $(y_1, \ldots, y_n)$, we can estimate $\mathbb{E}[Y|X]$ using the Nadaraya–Watson estimator $\hat{m}$, given by Equation 3.

$$\hat{m(z)} = \frac{\sum_{i=1}^{n} K\left(\frac{z-x_i}{h}\right) y_i}{\sum_{i=1}^{n} K\left(\frac{z-x_i}{h}\right)} \tag{3}$$

We created Figure 3 using this technique. Additionally, we used kernel regression to identify transformations of our key variables to ensure that our predictors were linearly related to our response. The transformations that we arrived at are as follows:

$$\begin{aligned} \Delta U_{k,j} &\mapsto \log\left(U_{k,j} + \left|\min_j U_{k,j} + 1\right|\right) \\ F_j &\mapsto \log(F_j) \\ D_j &\mapsto \log(D_j + 1) \end{aligned} \tag{4}$$

The relationship between individual transformed predictors and the transformed response are shown in Figures 5.

3.3 Linear Regression

The final method of analysis used in this project was linear regression. Compared to kernel regression and more complex machine learning models, linear regression is a very simplistic modeling technique that assumes that our response is a linear combination of our predictors. Although this model may be simplistic, we chose to use a linear model rather than a more complex algorithm in order to have clear interpretations of the coefficients that our model outputs. This is crucial for this project, where we are aiming to assess and interpret the relationship between our predictors and the change in unemployment rate.

In total, we ran three linear models. Model 1 and Model 2 include the raw variables described in Section 2, while Model 3 utilizes the transformed variables denoted by Equation 4. All of these models were implemented in R using the `lm` function (see Section **??**). The final results of our linear models will be covered in Section 4.

4 Results

The main results from our project are the regression outputs for our various linear models. The outputs of the linear models can be found in Table 2. In summary, we find that COVID case numbers are completely insignificant in predicting changes in the unemployment rate, while COVID death tolls are significant in this regard. Additionally, it appears as though rural lending percentage is the only predictor that is negatively correlated with $\Delta U_{2020,j}$. Conversely, it appears as though the amount of forgiven PPP lending has a positive correlation with changes in the unemployment rate, which seemingly contradicts their intended purpose.

Table 2: Regression Results

Variable	**Model 1**	**Model 2**	**Model 3**
Intercept	3.75***	4.05***	1.08***
(std. err.)	(0.08)	(0.08)	(0.14)
F_j	0.50***	0.51***	0.08***
	(0.13)	(0.13)	(0.004)
R_j	-1.66***	-1.52***	-0.06***
	(0.10)	(0.10)	(0.02)
C_j	0.05	0.04	
	(0.04)	(0.04)	
D_j			0.84***
			(0.14)
$\Delta U_{2019,j}$		0.46***	0.63***
		(0.04)	(0.04)
N	3088	3088	3088
R^2	0.169	0.208	0.306

$^{***}p < 0.01$, $^{**}p < 0.05$, $^{*}p < 0.1$

Based on the R^2 values for the three models, we can see that the transformed model had the best fit. Given this information, our discussion in Section 5 will focus specifically on Model 3. Additionally, the residual plots from Model 3 are given in Figure 6.

5 Discussion

5.1 Forgiven Loan Amount

As seen in Table 2, the variable F_j is positively associated with $\Delta U_{2020,j}$. This result seems to be counterintuitive. At first glance, one might conclude that the PPP actually contributed to rising unemployment, rather than combating it. However, we believe that the true story is more complicated than this. Referring back to Table 2, we can see that D_j is also positively associated with $\Delta U_{2020,j}$, indicating that as COVID deaths increase, so does unemployment. Now, consider the case of a county with a rising number of COVID deaths. As more workers succumb to the disease, employers will have more incentives to take out PPP loans to keep their businesses alive despite the shrinking workforce. Thus, we end up with a sort of confounding effect where the increase in COVID deaths contributes to both an increase in unemployment and an increase in PPP lending.

5.2 Rural Lending Percentage

Out of all of the factors in Table 2, only R_j is negatively associated with $\Delta U_{2020,j}$. If we take the rural lending percentage to be a measure of how rural a county is, this suggests that more rural counties were able to avoid rises in unemployment than urban counties. There are many possible explanations for this phenomenon. First, we could point to the low population density in rural regions, which would make it harder for COVID to spread. At the same time, this association might be due to the types of banks located in rural counties. Specifically, we note that community banks are mostly located in rural counties. These community banks usually serve a smaller population of businesses and are more "rooted" in the communities they serve when compared to large, corporate banks. This may have had an effect on the ability of rural businesses to receive PPP loans in a more timely manner. One final explanation that we present is that the types of industries located in rural counties are more resilient to volatility in the labor market. Take, for example, agricultural industries. Agriculture is an industry that is necessary to human existence and, thus, very difficult to fully shut down. This is especially true during periods like the COVID-19 pandemic when many countries and corporations halted shipping activities. This unique situation may give the agricultural industry a strong resilience towards shocks in the labor market.

5.3 Future Steps

We acknowledge that the results in this paper are limited. Our linear models were not able to explain much of the variance within our dataset (see Figure 6). Thus, we believe that future work is needed to fully uncover the effectiveness of PPP loans in mitigating unemployment. In particular, we believe that future research should include a wider variety of predictors in order to explain more of the variance in our data. Some recommendations for these variables include, but are not limited to:

- Bank Size
- Industry Type
- Population Density.

Additionally, we believe that the data used for this project is well suited to other types of statisical models. In particular, since our data includes timestamps, we believe that time-series analysis could be used to forcast the unemployment rate in a given month given a set of predictors. Another possibility is to use an econometrics model such as Panel Regression. This model seems appropriate given that our data contains both a temporal component and a spatial one.

6 Appendix A: Figures

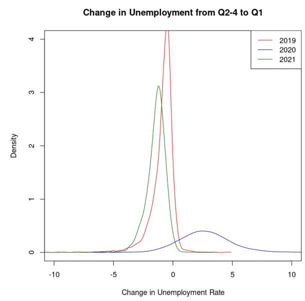


Figure 1: $\Delta U_{k,j}$ across counties

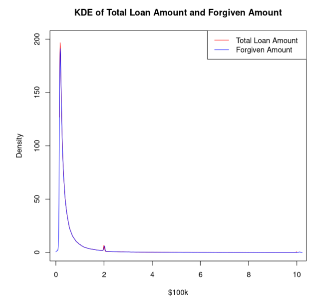


Figure 2: Total Lending vs Forgiven Lending

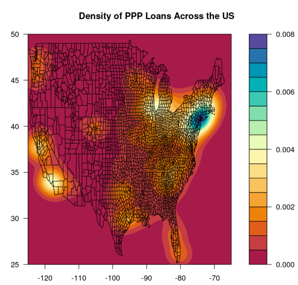


Figure 3: 2D KDE of PPP Loan Distribution

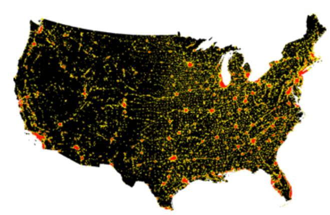
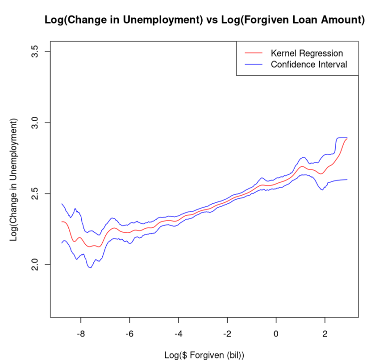
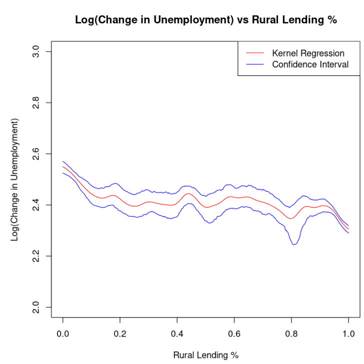


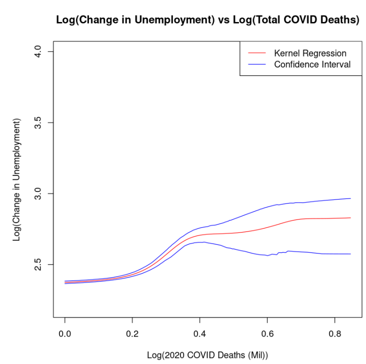
Figure 4: US Urban Population



(a) $\Delta U_{2020,j}$ vs F_j

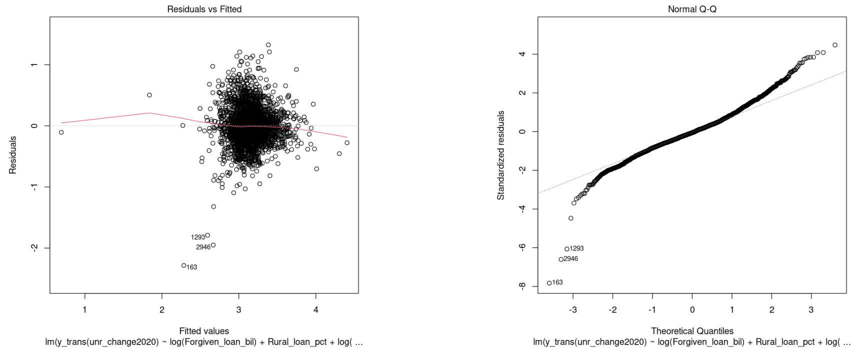


(b) $\Delta U_{2020,j}$ vs R_j



(c) $\Delta U_{2020,j}$ vs D_j

Figure 5: Transformed predictors vs transformed response.



(a) Residual vs Fitted Plot

(b) Normal QQ Plot

Figure 6: Residual Plots for Model 3

7 Appendix B: Code

7.1 Python Code

```
def add_county(data):

    import pandas as pd

    counties = pd.read_csv("./ZIP-COUNTY-FIPS_2017-06.csv", dtype='object')

    data['BorrowerZip'] = data['BorrowerZip'].str[:5]

    data = data.merge(counties, left_on = ['BorrowerState','BorrowerZip'], right_on = ['
        STATE','ZIP'])

    drop_cols = ['ZIP','STATE', 'COUNTYNAME','CLASSFP','Gender','Veteran',
                 'NonProfit','Race','Ethnicity','UTILITIES_PROCEED','RENT_PROCEED',
                 'PAYROLL_PROCEED','REFINANCE_EIDL_PROCEED','HEALTH_CARE_PROCEED','
                     DEBT_INTEREST_PROCEED',
                 'MORTGAGE_INTEREST_PROCEED','FranchiseName','ProcessingMethod']

    data.drop(drop_cols,axis=1,inplace=True)

    data.rename(columns={'STCOUNTYFP': 'BorrowerFIPS'},inplace=True)

    return data


def load_PPP(file_path):

    import pandas as pd
    import numpy as np

    data = pd.read_csv(file_path)

    data = data.loc[~(data['BorrowerCity'].isin(['','N/A',np.nan])) &
                    ~(data['BorrowerState'].isin(['','N/A',np.nan,'AK','HI'])) &
                    ~(data['BorrowerAddress'].isin(['','N/A',np.nan])),]

    data = add_county(data)

    data['full_address'] = data['BorrowerAddress'] + ',' + data['BorrowerCity'] + ',' + data
        ['BorrowerState']

    return data

```

```python
def main():

    import argparse

    parser = argparse.ArgumentParser()

    parser.add_argument("-f", "--file_name",help="file to load",required=True)
    parser.add_argument('-o','--output_name',help='output file name',required=True)

    args = parser.parse_args()

    PPP_loans = load_PPP(args.file_name)

    PPP_loans.to_csv(f'{args.output_name}.csv', index=False)


if __name__ == "__main__":

    main()
```

load_PPP.py

```python
def load_covid(file):

    import pandas as pd

    data = pd.read_csv(file, dtype={"fips":str})

    data = data.loc[(data['date'].str.startswith("2020")) & (~data['fips'].isna()),['date','fips','cases','deaths']]

    cases = data.groupby('fips').sum().reset_index()

    return cases

def main():

    import argparse

    parser = argparse.ArgumentParser()

    parser.add_argument("-f", "--file_name",help="file to load",required=True)
    parser.add_argument('-o','--output_name',help='output file name',required=True)

    args = parser.parse_args()

    cases = load_covid(args.file_name)

    cases.to_csv(f'{args.output_name}.csv', index=False)

if __name__ == "__main__":

    main()
```

load_covid.py

```python
def load_unemployment(file):

    import pandas as pd

    data = pd.read_csv(file)

    results = pd.DataFrame()

    results['BorrowerFIPS'] = data['Region Code']

    for i in range(2019,2022):
```

```python
        results[f'{i}_q1'] = data[[f'{i}-{j:02d}-01' for j in range(1,4)]].mean(axis = 1)

        results[f'{i}_q2'] = data[[f'{i}-{j:02d}-01' for j in range(4,7)]].mean(axis = 1)

        results[f'{i}_q3'] = data[[f'{i}-{j:02d}-01' for j in range(7,10)]].mean(axis = 1)

        results[f'{i}_q4'] = data[[f'{i}-{j:02d}-01' for j in range(10,13)]].mean(axis = 1)

        results[f'{i}_q234_avg'] = results[[f'{i}_q{j}' for j in range(2,5)]].mean(axis = 1)

        results[f'{i}_q234-q1'] = results[f'{i}_q234_avg'] - results[f'{i}_q1']

    return results

def main():

    import argparse

    parser = argparse.ArgumentParser()

    parser.add_argument("-f", "--file_name",help="file to load",required=True)
    parser.add_argument('-o','--output_name',help='output file name',required=True)

    args = parser.parse_args()

    unr = load_unemployment(args.file_name)

    unr.to_csv(f'{args.output_name}.csv', index=False)

if __name__ == "__main__":

    main()
```

load_unemployment.py

7.2 R Code

7.3 C Code

```c
#include <math.h>

#define M_PI 3.14159265358979323846

void KDE(int* n, int* m, double* x, double* g, double* y, double* bw){

    /*
    n - length of observed value vector
    m - length of grid vector
    x - vector of observed values
    g - grid of points to estimate
    y - result vector
    */

   for (int i = 0; i < *m; i++)
   {

    double sum = 0.0;

    for (int j = 0; j < *n; j++)
    {
        sum += (1 / (*bw * sqrt(2*M_PI))) * exp(-(x[j] - g[i]) * (x[j] - g[i]) / (2 * *bw * *bw));
    }

    y[i] = sum / (*n * *bw);
```

```c
    }

}
```

kde.c

```c
#include <R.h>
#include <Rmath.h>

void NW_estimate(double *x, double *y, int *n, double *b,double *g, int *m, double *est){

    int i,j;
    double a1,a2,c;

    for(i = 0; i < *m; i++){

        a1 = 0.0;
        a2 = 0.0;

        for(j=0; j < *n; j++){

            c = dnorm((x[j]-g[i])/ *b,0,1,0);
            a1 += y[j] * c;
            a2 += c;
        }

        est[i] = a1/a2;

    }


}
```

kre.c

References

[1] Ayşegül Şahin, Murat Tasci, and Jin Yan. "The unemployment cost of COVID-19: How high and how long?" In: *Economic Commentary (Federal Reserve Bank of Cleveland)* (2020), pp. 1–7. DOI: 10.26509/frbc-ec-202009.