Use Big Data to Predict Unemployment During the COVID-19 Pandemic

Jamie Chen(✉)

Shenzhen Foreign Languages School, Shenzhen, China
chenzhixiang4947@163.com

Abstract. Since the pandemic, unemployment has always been a concern in the United States. The aim of the paper is to use big data to predict unemployment in the U.S. Based on results from correlation matrix and OLS regression, I conclude that the Google Trend Index is a helpful reference to understand the situation of unemployment rate in every state in the United States. This paper contributes to the literature by connecting unemployment and Google Trend index at the state level, as well as providing a deep understanding of predicting economic factors using internet real-time big data.

Keywords: Big Data · Unemployment · Google Trend Index · COVID-19

1 Introduction

From 2020, economies across the world are feeling the negative effects of nationwide lockdowns and economic uncertainty as a result of the coronavirus pandemic [Amburgey, Aaron, Serdar Birinci 2020]. The United States is deeply affected by a severe pandemic, followed by a sharp rise in unemployment. Some states in the United States are forcing the closure of public places and non-essential stores, while others are closing on their own. While businesses in many sectors are experiencing losses, the earlier negative effects of the coronavirus pandemic have especially affected the service sector. Businesses in the service sector have been the first to shut down, and workers in this sector have been more likely to experience layoffs early in the crisis [Amburgey, Aaron, Serdar Birinci 2020].

The U.S. labor force participation rate during the pandemic was significantly lower than the period before the pandemic. According to CNBC news [Amburgey, Aaron, Serdar Birinci 2020], the September jobs report hints at unemployment benefits' muted role in the pandemic labor market. Figure 1 shows the U.S. job growth rate from mid-2020 to mid-2021. At the beginning of the pandemic, the job growth rate was decreasing significantly.

In recent years, the use of big data and advanced computing techniques has made people possible to nowcast the present unemployment rate and forecast the future unemployment. Google Trends data is intended as real-time data. Google trend data allows us to track the specifics of unemployment in the United States and it can show us daily and annual search leaderboards during the pandemic.

G. Guan et al. (Eds.): ICBDSS 2022, AHCS 8, pp. 545–552, 2023.
https://doi.org/10.2991/978-94-6463-064-0_56

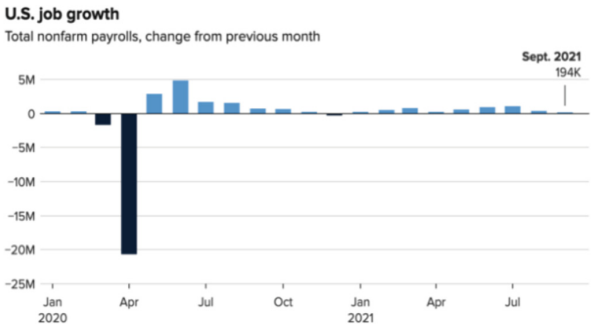

Fig. 1. U.S. job growth in 2020 and 2021 (Source: Google)

In this paper, I'm going to focus on the unemployment rate in various U.S. states and see whether big data can accurately predict unemployment claims or not. First, I will look at the impact of COVID-19 on unemployment rates in states across the United States. Then, I will relate unemployment data to Google Trends data to show some correlation matrix and OLS regression tables, which will help us understand the relationships between unemployment and Google Trends. The remainder of this paper proceeds as follows. I introduce my data in Sect. 2. I will then describe my results in Sect. 3. I will describe my conclusion in Sect. 4.

2 Data

2.1 Google Trend Data

Google Trends is a Google analytics tool based on search data. This big data technique provides a time series index that reflects the volume of queries of those users who have introduced the associated keywords into Google Search in a specific zone. Askitas and Zimmermann (2009) used search keywords "*job search*", "*labor office*", and "*short-term work*" to evaluate prediction performance. In addition, Google Trends can also provide data trends in different regions at the same time as well as different times in the same region. The query index is calculated by dividing the total number of search queries for a keyword in a specific area by the total number of search queries in the area in a certain period. A normalization to 100 is made for the maximum query share in that period, and a normalization to 0 is made for the query share at the initial time (Choi and Varian, 2012) [Simionescu, Mihaela, Zimmermann 2017]. The major advantage consists of the ability to define a group of relevant variables and build the associated content based on the definition and merge for keywords. Therefore, the effects of different concepts can be easily analyzed. In this paper, since I want to study whether Google Trends can accurately predict the change in the unemployment rate at state-level, I search "unemployment rate," "unemployment subsidy," "job search" and other related keywords in Google Trends to obtain data about unemployment in the United States.

2.2 Unemployment Insurance Initial Claims

The unemployment insurance weekly claims data is from the United States Department of Labor. The increase in layoffs in the U.S. has translated into a dramatic rise in unemployment insurance (UI) claims [Amburgey, Aaron, Serdar Birinci 2020]. Unemployment insurance (UI) claims are a good predictor for real-time unemployment rates. They can be used to forecast near-future unemployment for local areas as well as industrial sectors. UI claims provide a quicker and more flexible read on the market that potentially benefits policymakers and economic researchers [Zheng, Claire. 2020]. On March 21, 2020, the number of weekly U.S. UI claims reached nearly 3.3 million, its highest level ever. In this paper, I will treat it as the dependent variable for providing information about unemployment in the United States.

2.3 Other Data

The New York Times includes state-level COVID confirmed cases and deaths as unobserved state-level characteristics controlled in OLS regression at the state level. For example, I add detailed data such as COVID cases or deaths to my OLS regression in table.

3 Big Data Analysis

3.1 Google Trends Data Chart

First, I search for four words on Google Trends and observe the changes in the number of searches for these four words in different states in the United States. I use a kind of chart, and its y-axis represents search volume and ranges from 0 to 100. “0” means the search volume is 0 at that time, while “100” means the maximum search volume. The words are “unemployment”, “unemployment insurance”, “unemployment benefits” and “unemployment claim”. According to the charts, we can see that the situation is not the same in all states of the United States. South Dakota, for example, is one of the safest states relative to other states. Its four curves do not exceed 30 and are relatively stable. The four curves peaked in Nevada, Texas, Alabama and Hawaii separately.

Then let me introduce some interesting findings in the chart. Why is Kentucky’s unemployment insurance so large in variance? In Kentucky, unemployment insurance premiums have risen from $8.50 per month in March to $13.50 per month in November. According to the Kentucky Insurance Department, that’s a 33% increase in the past year. In a statement, the agency said, “Since October, the unemployment insurance rate in Kentucky has risen from 9.4% to 10.4%.” The increase is not expected to be reserved. The insurance industry says the increases are a result of higher premiums, as well as new problems with the program, such as a lack of reliable claims data [Roberts, Brandon. 2020]. Over the course of several months beginning in late March, thousands of Kentuckians filed claims, and while new claims have recently receded, real unemployment remains at or above its highest level since the Great Recession [Charlton, John. 2021].

Then, why are searches for unemployment claims so low in South Dakota? Santos said South Dakota’s economy had fared “rather well” during the economic expansion.

Fig. 2. State-level Google Search Index

Prior to COVID-19, unemployment levels were consistently low. During the seven-week period from March 14 to May 2, a total of 37,645 South Dakotans filed initial claims for unemployment benefits, representing about 8.1% of the March labor force, according to Department of Labor data released. It's the lowest rate of initial jobless claims among all 50 states. The state also has a robust financial and community banking industry, spurred in part by usury law changes that eliminated the caps on interest rates and attracted Citibank and other credit card issuers [Wallace, Alicia. 2020]. In addition, Texas' unemployment benefits search is very high during the pandemic. According to research, Texas lost 1.3 million jobs in April 2020, as payroll employment fell at a historic rate of nearly 11.0%, and the unemployment rate climbed to 13.5%. In Texas, the median replacement rate for jobless benefits through Federal Pandemic Unemployment Compensation increased to 153%, significantly higher than the average replacement rate of 52% through the regular state unemployment insurance program. Nearly 72% of the 1.12 million Texans who claimed jobless benefits in the week ending Sept. 26 received benefits through the regular state program; the Pandemic Unemployment Assistance program accounted for 26% of total claimants, and the Pandemic Emergency Unemployment Compensation program accounted for 3% [Kumar, Anil. 2020] (Fig. 2).

3.2 Correlation Matrix

Second, I use correlation matrix to further explain whether Google Trends can accurately predict unemployment in United States or not. In addition, a correlation matrix describes correlation among M variables. Based on the matrix, we can find the relationship between the two variables. For example, the relationship between initial claims and job losses is negative. As we all know, when someone lose their job, they will usually apply for unemployment benefits, which means that the search volume of initial claims should increase. Other keywords like "job loss" also have a negative relationship with initial claims. Because when the economy grows, so do people's incomes. What's more, the country's unemployment rate will decrease during this time, which means

Table 1. Pairwise correlations

Variables	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Unemployment Initial claims	1.000						
(2) GT: income	0.077***	1.000					
(3) GT: income loss	-0.219***	-0.021	1.000				
(4) GT: job	-0.098***	0.004	0.018	1.000			
(5) GT: job loss	-0.280***	-0.038	0.332***	0.037	1.000		
(6) GT: labour	-0.190***	0.075***	0.246***	0.035	0.237***	1.000	
(7) GT: unemployment	-0.023	0.016	-0.049	0.083***	-0.043	-0.016	1.000

*** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

that fewer people will apply for unemployment benefits. Therefore, the search volume of initial claims will decrease. However, if people's incomes decrease, the effect will be totally different. The decline in people's incomes indicates that the economy is in recession. The unemployment rate will therefore rise. More and more people will apply for unemployment benefits. So the search volume of initial claims will then increase.

3.3 OLS Results

Unemployment initial claim is the dependent variable in Table 1. There are three independent variable groups in this table. In "only unemployment GT", all variables have a very strong relationship with initial claims since they are at a 0.01 significance level. In addition, the magnitudes of the variables are also very large, which enhances the strong relationship between these four variables and initial claims, especially the magnitude of unemployment benefits. The relationship between unemployment and initial claims is negative. For example, if the applications for unemployment benefits increases, the economy will go into recession, which means that more people will become unemployed.

The next column is "Add Other GT". It has several more independent variables than "only unemployment GT", such as income loss, job, job loss, and labour. Income loss and labour are at 0.05 significance levels, and the job is at a 0.01 significance level. So, they both have a strong relationship with initial claims. Next, I will explain why the remaining three new variables have a positive or negative relationship with the initial claims. First, income loss has a positive relationship with initial claims. Imagine that your income is decreasing, which means that you are probably facing unemployment. You will then apply for unemployment benefits. Therefore, the search volume of initial claims will increase. Second, the job has a negative relationship with initial claims. The relationship is very strong. Therefore, they will apply for unemployment benefits to support their lives. The search volume of initial claims will then increase. Third, both the job and the job loss have the same reason why they have a negative and positive relationship with

Table 2. OLS Regression

VARIABLES	(1)	(2)	(3)
	only unemployment GT	Add Other GT	Add COVID cases/deaths
GT: unemployment	-548.4***	-1,011***	-779.6**
	(84.00)	(375.2)	(349.0)
GT: unemployment benefits	824.0***	1,342***	665.6**
	(78.42)	(314.9)	(310.6)
GT: unemployment claim	661.0***	3,079***	2,604***
	(93.41)	(729.7)	(698.3)
GT: unemployment insurance	134.0***	1,115**	1,111**
	(49.95)	(484.3)	(477.0)
GT: income loss		506.3**	328.9
		(214.7)	(212.3)
GT: job		-2,044***	-1,251**
		(527.1)	(485.6)
GT: job loss		239.1	425.4*
		(227.1)	(230.0)
GT: labour		621.9**	413.1
		(268.2)	(259.2)
Cases/capita			-15.26**
			(6.919)
Deaths/capita			-176.3
			(406.0)
Constant	-31,450**	-198,270***	-52,065
	(12,199)	(74,605)	(77,247)
Observations	2,021	426	426
R-squared	0.483	0.467	0.518

Robust standard errors in parentheses *** $p < 0.01$, ** $p < 0.05$, * $p < 0.1$

initial claims separately. What's more, after we add some other independent variables, we can see that the relationship between initial claims and the previous four variables has become stronger (Table 2).

The last column is "Add COVID cases/deaths". It contains all the variables from "Add Other GT", plus cases precap and deaths precap. According to the chart, the size of the

original eight variables has changed again, which means that the relationship between initial claims and unemployment benefits also decreases. More importantly, since the significance of unemployment benefits has decreased, its relationship to initial claims is also not as significant as it is in "Add other GT". In addition, according to the chart, the significance of unemployment claims doesn't change. The magnitude of unemployment claims decreases. Therefore, the relationship between initial claims and unemployment claims decreases. Furthermore, searches for unemployment insurance barely changed, meaning that after adding some new independent variables, searches for unemployment insurance were unaffected. Moreover, income loss has the total opposite effect compared with unemployment insurance. We can also see that after adding the new independent variables, the relationship between initial claims and income loss drops sharply. Next, let's talk about the relationship between the job and the initial claims. According to the chart, both the magnitude of the job and the significance of the job have decreased. Obviously, this is also an example that is affected by the new independent variables. The relationship between them is not as significant as before. Finally, I will discuss labour. According to the chart, the magnitude of labour decreases from 621.9 to 413.1. What's more, it becomes insignificant after the new independent variables are added. Therefore, it's easy to conclude that it has almost no relationship with initial claims.

4 Conclusion

In recent years, unemployment has always been a concern for every country. In this paper, I choose the United States as the subject of my paper. The availability of real-time big data allows me to explain whether Google Trends can accurately predict the unemployment rate in the United States. I use Google Trends data, the Correlation Matrix, and OLS to support my viewpoint. Finally, I conclude that some correlation data is in line with expectations. We can use Google Trends as a reference to help us understand the unemployment rate in every state in the United States. Typically, real-time big data will provide predictions and insights for policymakers to reduce the severity of unemployment.

References

Amburgey, Aaron, and Serdar Birinci. 2020. "The Effects of Covid-19 on Unemployment Insurance Claims." Economic Synopses 9.

Askitas, Nikos, and Klaus F. Zimmermann. 2009. "Googlemetrie Und Arbeitsmarkt." Wirtschaftsdienst 89 (7): 489–96.

Charlton, John. 2021. "Troubled Unemployment System Still Problematic for Kentuckians | Whas11.Com." Accessed March 28, 2022. https://www.whas11.com/article/news/investigations/focus/one-year-of-kentucky-unemployment-during-covid-pandemic/417-771953b6-1aeb-4f64-86c2-0e40b7c6686c

Kumar, Anil. 2020. "Pandemic Unemployment Benefits Provided Much-Needed Fiscal Support." Southwest Economy, no. Fourth Quarter.

Roberts, Brandon. 2020. "Problems with Kentucky's Unemployment Insurance System." Accessed March 28, 2022. https://spectrumnews1.com/ky/louisville/news/2020/12/14/kentucky-unemployment-insurance-problems.

Simionescu, Mihaela, and Klaus F. Zimmermann. 2017. "Big Data and Unemployment Analysis." GLO Discussion Paper.

Wallace, Alicia. 2020. "Why Unemployment Claims Are so Low in South Dakota, Utah and Nebraska - CNN." https://edition.cnn.com/2020/05/12/business/states-lowest-initial-unemployment-claims/index.html.

Zheng, Ping Claire. 2020. "Predicting Unemployment from Unemployment Insurance Claims." Indiana Business Review 95 (2): 1–9.