Kevin Lu

Stat 431

Final Paper

Statistical Review of Hakan Yilmazkuday's "COVID-19 Effects on the U.S. Unemployment: Nationwide and State-Level Evidence from Google Trends"

Goal of the Study / Summary

The goal of the "COVID-19 Effects on the U.S. Unemployment: Nationwide and State-Level Evidence from Google Trends" by author Hakan Yilmazkuday is to investigate the relationship between nationwide Google search queries of "covid-19" and "unemployment" using daily data from the United States. The investigators also related the traffic of queries of "interest rate" to later traffic of "unemployment", suggesting that the Federal Reserve has a part to play in the redistribution of relief funds across states. The study aims to provide evidence for the effects of covid-19 on unemployment, on both a nationwide and a state-level basis.

Data/Stats

Data: The data used in this paper is observational data, as it is aggregated from daily queries of "covid19", "unemployment", and "interest rate", broken down by state, and collected via Google Trends. The data was collected between the dates of January 1st, 2020 and April 11th, 2020. The data is a record of the number of search queries for each of the 4 terms "covid-19", "unemployment", "interest rate", and "inflation". Additionally, a 5th piece of information is attached to "unemployment", which is the state in which the query was launched.

Bias: The shape of the Google Trends interest graphs for the terms “covid-19” and “unemployment” for example show that they have alternated in short intervals between peak interest and half of peak interest, coinciding with times of major news such as press conferences from world leaders as well as official statements by medical organizations. A day with any sort of large announcement regarding coronavirus or the economy appears to have an increase in queries left in its wake. For example, for “covid-19”, the first time it peaks is March 16, one day after the CDC recommended limiting the size of gatherings. Another peak came with the news that Prime Minister Boris Johnson had locked down Britain. Another peak on March 28 trails behind news of President Trump signing a stimulus package, as well as on the 28th news of a CDC travel advisory to New Yorkers. In other words, the interest in the search terms is not evenly spread out, but clustered around dates where the public received relevant and big news releases.

Another factor that accounts for possible bias is industry. Certain industries are impacted much more heavily than others, and this can be reflected in the specializations of different states. For example, California has Silicon Valley representing much of America’s tech industry, whereas Wisconsin mainly focuses on agriculture. Tech jobs can oftentimes be done remotely so the crisis affects ordinary workers a bit less in comparison to other blue-collar workers in agriculture and such that may not have that luxury. The necessity of social distancing and stay-at-home orders means that a variety of problems arise such as a break in the supply chain when delivering goods, as well as layoffs as companies struggle. Thus, we can expect some states with industries more affected by the crisis to be more likely to have higher filings for unemployment.

In addition, a factor not mentioned is government policy. Each state has approached the coronavirus crisis differently, whether it be total lockdown or recommended social distancing and limited party size. In addition, the speed of reaction to the coronavirus also heavily affects how severe the impact is in the state, and these factors all play into how people query online.

Underlying Assumptions: The investigation is founded upon certain assumptions that grant the study meaning, as well as justify the choice of model and methods. One of the assumptions is that the interest via the queries is in some way able to gauge the level of impact or interest of the said search term. By this I mean, for example, the real-life effect implied by a large number of “unemployment” queries in a state. The study is assuming here that this translates into a lot of people with the intent to file for unemployment. This relationship is logical, but difficult to quantify as there isn’t such a conversion rate known to us.

Methods:

Structural Vector Autoregression

Vector Autoregressions are statistical models used “to capture the linear interdependencies among multiple time series” according to Wikipedia. The model used in this study is called a Structural Vector Autoregression (SVAR), which is a type of VAR that possesses a diagonalized covariance matrix for its error terms (structural shocks). They make explicit identifying assumptions which isolate parameters’ behavior and their effects while minimizing reliance on assumptions of a variable’s behavior.

Partial derivatives are taken on each variable while keeping the others constant. This thus generates n + 1 differential equations, one for each variable and then one. Each differential equation, differentiated with respect to one variable, is represented as a vector. The n+1 vectors can then be arranged into a matrix and solved. Solving the matrix will then optimize the model. SVARs are appropriate for this study as the error terms are of interest in this study when finding the covariance between the query terms.

Bayes Estimator

The Bayesian approach refers to the use of the Bayes Estimator which, according to Wikipedia, minimizes posterior expected value of a loss function or maximizes posterior expected value of a utility function. The Bayes estimator acts as a function of observable variables, and estimates the value of an unknown prior distribution. The goal is to find an estimate with the lowest mean squared error. This is commonly done in Bayesian statistics when modelling, and so it appears justified to apply this technique to the SVAR with the independent normal-Wishart priors as done in this investigation. The medians of each distribution were considered as Bayesian estimators, and the quantiles of the Bayesian estimators were used to construct credible intervals which appear in the tables to serve as evidence for varying impact between states.

Independent normal-Wishart priors

Priors are the probability distributions expressing some belief in a quantity before some evidence is taken into account. The normal-Wishart distribution is a multivariate family of continuous probability distributions with four parameters. The usage of this distribution is indeed appropriate for this study as it did indeed involve four independent parameters. The usage of the distribution is also highly appropriate due to the fact that it is the distribution for a sum of squares. What we are interested in when working with SVARs is minimizing the sum of squared error, so this distribution will be expected to be suitable for the data and model at hand.

Credible Intervals

Credible Intervals are analogous to confidence intervals in that they are defined for the median very similarly to how confidence intervals are defined for the mean. Credible intervals are Bayesian in nature while confidence intervals are frequentist. According to Wikipedia,

> “Bayesian intervals treat their bounds as fixed and the estimated parameter as a random variable, whereas frequentist confidence intervals treat their bounds as random variables and the parameter as a fixed value. Also, Bayesian credible intervals use (and indeed, require) knowledge of the situation-specific prior distribution, while the frequentist confidence intervals do not.”

Bayesian statistics has an advantage over frequentist statistics in our case, as it lacks the requirement of knowing exactly what distribution to use for the data. The credible interval in Bayesian statistics is simply the range containing a particular percentage of probable values. The range is not fixed. The conclusion drawn from a credible interval is much more easily interpretable than that of a confidence interval; in fact, a common mistake when discussing results is actually attributing the credible interval interpretation to a confidence interval.

In frequentist statistics, the confidence interval is constructed via a given confidence level from an infinite number of independent samples. The confidence level represents the proportion of possible confidence intervals that contain the true value of the unknown population parameter. The confidence interval is random as it depends on random samples.

Credible intervals incorporate problem-specific contextual information from the prior distribution whereas confidence intervals are based only on the data. A prior distribution of an uncertain quantity is the probability distribution that would express one’s beliefs about this quantity before some evidence is taken into account.

Visualization:

- Table displaying medians and credible intervals of cumulative national effects
  - The usage of medians and credible intervals is in line with the Bayesian approach taken to model the data, so it is appropriate as a measure of central tendency. The table effectively communicates the predictive effects of each of the four search terms on themselves and the others on a state-level nationwide, which allows us to see the disparity in magnitude for the impact on states. This serves as fairly compelling evidence in favor that there is an association between the queries and future queries, and the difference between states indicates that the impact of the current crisis is experienced differently between states too.
- The Appendix has a table of summary long-run “unemployment” effects on each state, detailing results of the search terms “covid-19”, “unemployment”, “inflation”, and “interest rate”.
  - This table shows the median long run effects of each of the search terms by each state. This breakdown by state makes it clear how different search terms affect different states differently. For example, it is clear from the table that “covid-19” affected Alabama to a much lesser degree than Arizona (2.503 to 6.972). This does make sense as Arizona was one of the first places in the United States to report cases of covid-19, so residents would naturally be more wary of the virus and search for it more. In another example, “unemployment” affects states in radically different ways, like with Florida (3.554) and Hawaii (-3.656). Differences like this are made easily visible such that it is extremely compelling when trying to convince oneself that there is indeed a difference in effects across states.

Conclusion

To reiterate the purpose of the paper, the authors investigated the relationship between Google queries of "covid19" and "unemployment" using daily search data nationwide, by including relevant search terms into a structural vector autoregression model. The author approached this study with the knowledge that historical decomposition analysis indicates a belief that the spike in “unemployment” queries is largely explained by the queries of “covid-19” in mid-March.

The data drove a conclusion that the traffic of searches for “covid-19” can predict later search traffic for “unemployment”, with “interest rate” providing little predictive power and “inflation” almost none. Another conclusion drawn as part of a side investigation was that traffic for “interest rate” was also able to predict traffic for “unemployment”. More importantly, both of these conclusions also supported an additional insight that the impact of the coronavirus crisis had different magnitudes across different states and that different states had varying levels of need for federal assistance, which makes sense as different states specialize in different industries which are affected unevenly by the crisis and government response.

For example, Florida is a popular travel destination and retirement state, and so the danger of coronavirus there was significantly higher, leading to harsher economic consequences. States like Michigan focusing on manufacturing have seen many factories temporarily closed, and agricultural states like Wisconsin see problems with the supply chain in getting produce shipped on time with orders being cancelled and delivery services being impacted. In places like California’s Silicon Valley, where the tech industry lives, ordinary workers tend to be less affected as working from home is an option, and so the economic impact would be to a lesser degree as fewer people are at risk for unemployment.

Results from accumulating impulse responses indicate that an increase of one unit of positive nationwide “covid-19” queries yields an equivalent increase of 8.1 units in nationwide “unemployment” queries after 2 months. This actually ranges over 2.5 - 9.3 once broken down into state-level parameters, indicating that the search query quantity varies greatly by state, which suggests unequal effects of “covid-19” across states.

Another conclusion drawn is associated with the Federal Reserve’s reaction to the crisis, which shows that one unit in negative national “interest rate” shock yields 0.4 units of corresponding accumulated reduction in nationwide “unemployment” after 2 months. The effects range between 1.3 and 0.5 units increase per unit negative national “interest rate” shock, based on an additional state breakdown. To interpret these numbers, the Federal Reserve’s expansionary monetary policy () appears to have aided in reducing nationwide unemployment. However, between states the effects vary with the same nationwide “interest rate” shock helping Colorado reduce “unemployment” by 1.28 units, while only helping South Dakota reduce it by 0.53 units. This suggests evidence for the national monetary policy affecting redistribution across the states.

Overall Impression

The paper was written in a relatively easy to understand and straightforward manner. The explanations were generally sufficient, and the tables and figures shown indicate extensive proof of the trends observed, displaying them in a manner that is easily comprehensible visually. The use of data was done in a statistically meaningful and clear manner, with assumptions stated and justified. The techniques chosen to model the data were appropriate and sufficient justification and explanation of those techniques were documented such that I did not discover any clear fault with them. He kept consistent with his use of Bayesian techniques and used them in a manner orthodox with other similar statistical literature. The author referenced numerous other works by other authors which supported his claims, most of which were written in the same year or just a few years before. Thus the author is able to claim with some degree of confidence that his work is able to serve as evidence as it further reinforces what has been published already. The approach of using Google query data is an interesting one that differentiates it from the literature already published, but with this approach comes a few limitations as assumptions are made that search frequency is representative of impact. Logically this is sound but with conversion factor being provided the numbers are difficult to translate into actionable benchmarks.

The authors ran the numbers and gave us the model, but relatively little was done in terms of suggesting what we can do with given results. I think that I would have liked to see a proposed course of action for the Federal Reserve after acknowledging and even proving the varying levels of need and impact of the virus outbreak amongst the states nationwide. Currently, from this study we know that the effect of “covid-19” on “unemployment” varies heavily by state, and that interest in “interest rate” also varies heavily by state. I believe that the authors could have provided the natural extension into showing how their model could be used to allocate relief funds across states to accurately facilitate need-based distribution.

This paper represents my own work in accordance with the Code of Academic Integrity

-Kevin Lu