Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021)

Causal Analysis of Impact Factors of COVID-19 in China

Wen-Xun Tang\textsuperscript{a}, Haifeng Li\textsuperscript{a*}, Mo Hai\textsuperscript{a}, Yuejin Zhang\textsuperscript{a}

\textsuperscript{a}School of Information, Central University of Finance and Economics, Beijing, 102206, China

Abstract

Mobility, group awareness, and temperature are considered as the important factors that may impact the increase in confirmed cases of the COVID-19\textsuperscript{[1]}. This paper aims to verify the above factors on the COVID-19 and show the possible confounding factors of each research variable in reality. Based on this, we collected data about the epidemic from January 20, 2020 to February 24, 2021, including the relevant data of 31 provinces and regions in China. Plus, we use the directed acyclic graph (DAG)\textsuperscript{[2]} to show the causal relationship between the above influencing factors and the confirmed daily epidemic cases, and the confounding is estimated based on DAG. The effective adjustment set of factors are used to perform the regression of the total causal effect among the explanatory variables and the confirmed cases of the epidemic using negative binomial regression.

Through the comprehensive causal analysis of the decisive factors for the COVID-19, we provide strong evidence for population mobility, group awareness and the impact of weather on the epidemic, and estimates the possible confounding factors in all aspects of society. Incorporating the above factors, we provide suggestions for future decisions on the prevention of large-scale epidemics.

© 2021 The Authors. Published by Elsevier B.V.
This is an open access article under the CC BY-NC-ND license (https://creativecommons.org/licenses/by-nc-nd/4.0)
Peer-review under responsibility of the scientific committee of the The 8th International Conference on Information Technology and Quantitative Management (ITQM 2020 & 2021)

Keywords: COVID-19, Causal analysis, Impact factor

1. Introduction

With the spread of the COVID-19 in China, various studies on the mechanism of the spread of the epidemic have also shown different results. In order to evaluate the impact of key factors on confirmed cases of the epidemic, the existence of confounding factors and the interaction with key factors are also considered. China has made outstanding contributions to epidemic prevention and control. All provinces have linked epidemic prevention and coordinated planning. Information on all aspects of society is transparent and open, it can be more convenient to get epidemic-related data. The data published by China provinces have consistent observational...
properties and a uniform tracking level of epidemic data, so it is more appropriate as a research direction.

We first briefly introduces and reviews the method of using causality diagrams as a technical tool to establish the knowledge background of the mechanism of transmission and infection of the COVID-19. Then, we collect data and integrate the observational data into a DAG. Finally, we use the model and calculation rules, and statistical methods to perform regression fitting to estimate causal effects.

2. Casual Model

We use the structured causality model (SCM) proposed by Judia-Pearl et al. to establish the spreading mechanism of the epidemic under the action of related factors, and estimate the causal effect between the driving factor and the response variable.

2.1. Cause and effect diagram

SCM consists of two parts, cause and effect diagram and corresponding structural equation[3]. The construction of the causality diagram is derived from the Bayesian network, so it uses the same set of conditional independence criteria.

Figure 1(a) represents a chain, which is expressed as X→Y→Z, in which X and Y are independently distributed under the Z condition. The second structure is a fork (Figure 1(b)), which is expressed as X←Z→Y, X and Y, again, are independently distributed under the Z condition. As shown in Figure 1(c), it is a collision and it is expressed as X→Z←Y. When Z is not a condition, X and Y are distributed independently.

The above principles of conditional independence based on Bayesian networks play a key role in determining the structural causal model and identifying causal effects.

2.2. Causal inference

The process of using causal identification algorithms to distinguish causal effects among random variables and statistical inference is called the causal inference. When estimating causal effects between variables, not all structures of the DAG are known or variables are observable, and potential variables will have confounding effects on the causal relationship between variables. Therefore, an important content of causal inference is to eliminate or control confounding variables and determine a clear causal relationship.

The backdoor criterion is an important method of causality identification, whose purpose is to eliminate the "false correlation" between data caused by confounding factors, and find the true causality. Besides, When the backdoor adjustment is not enough to meet the experimental requirements, the front door adjustment is needed.

3. R package

R is a programming language which has many packages for modeling and evaluating causal effects in causal inference. This paper mainly uses the pscl package and the digtty package for causal inference.

The main goal of the pscl package is to evaluate the structure of the overall causal model when the real causal structure has not yet been clarified, and give the effect of intervention under the evaluated causal structure.
model\textsuperscript{[4]}. In this experiment, the main task is to determine the minimum effective adjustment set when evaluating the causal effect of the research variable on the observed variable, eliminate the data bias caused by confounding factors, and obtain a categorical causal effect of the research variable on the observed variable.

The idea of the digtty package is derived from the DAGitty web application, which can access all the functions of the latter for statistical calculations, and provides new functions on this basis. It is more robust in causal inference than the pscl package. The difference is that pscl is limited to small causal models (such as the number of variables less than 10), while the digtty package can handle large data sets. It provides the function of transforming a given DAG into a matrix, and inferring the causal relationship between a variety of research variables and observation variables based on the DAG, and finally getting the adjustment set as the output result. It is based on the necessary and sufficient standards for causal models for different types of graphs proposed by Johannes Textor et al., the generalized adjustment criterion\textsuperscript{[6]}, to evaluate and output the adjustment set.

4. Data

The data of COVID-19 includes the number of daily cases at the provincial and regional levels updated daily on Tencent's real-time tracking website\textsuperscript{[7]}, including the cumulative number of confirmed cases, deaths, cured, new confirmed cases, deaths, cured.

The awareness describes the public awareness and cognition of the CPVID-19. Since emotional color is difficult to directly measure, this article uses Baidu Index website\textsuperscript{[8]} to search for "COVID-19" to reflect the group state of consciousness for the new crown. In order to assess population mobility, we uses data published by Baidu Qianx\textsuperscript{[9]}, which includes the migration, and is measured by the changes in the population mobile phone positioning between provinces or regions.

Temperature is considered to be an important factor affecting the spread of the COVID-19, but other papers show that the temperature factor is not critical\textsuperscript{[10]}. Therefore, when considering climatic factors, in addition to the temperature of each province and region, the wind and the air humidity are also included. The historical temperature and wind data sets are from the TianQiHouBao website\textsuperscript{[11]}, and the air humidity is obtained through the Wunderground and China AQI monitoring platform\textsuperscript{[1]}. In addition to the above factors that may have effect on the spread of the epidemic, other variables at the provincial and regional levels, including the annual GDP of each province and region, the level of emergency response, and the time to resume work and production are also considered. GDP measures the local economic development level. The population structure is collected and summarized through the Heihong population database\textsuperscript{[14]}. The population comes from the annual national economic and social development bulletins of various provinces and regions. Since the aging population is missing, we estimate the value from a sample survey of the National Bureau of Statistics\textsuperscript{[15]}.

5. Method

5.1. DAG and adjustment sets

This paper used a DAG as a tool to analyze the causal relationships between several exposures and COVID-19 spread(Figure 2). As can be seen, the research variables and the observed variables of epidemic cases are embedded in a DAG. Given the position of each variable node and the direction of the directed edge, a causal graph is obtained as a graphical representation of subsequent hypothetical causal inferences. Due to the complexity of the data variables, there is no direct evidence showing that mobility, group awareness, and weather have a direct causal link to the COVID-19. Here we only estimate the total causal effect of the research variables on the number of confirmed cases in the epidemic observation.
Figure 2: DAG of determinants of reported COVID-19 cases on the district level.

The adjustment set is a set of variables that estimate the effectiveness of causal effects. The built-in function of the R package encapsulates the backdoor adjustment algorithm for the cause and effect diagram. The experiment uses the R package dagitty to graphically represent the DAG and obtains the minimum adjustment set, and uses the pscl package to find the best adjustment set from the minimum adjustment set results. The output of the adjustment set may have multiple results. Under different adjustment sets, select the effective set with the highest pseudo-R2 value to obtain the causal effect under the best explanatory variables for the number of epidemic cases.

5.2. Negative binomial regression

As a basic counting regression model, the Poisson Regression (PR) model requires the mean and variance of the data to be equal, but it is difficult to achieve this result in the actual data fitting, so the problem of excessive dispersion often occurs. The negative binomial regression model is a popularization of the PR model, which can solve such problems and has been successfully applied in the field of epidemiology, becoming a common method to analyze the infection mechanism of epidemic cases.

The model assumes that there is a log-linear relationship between the expected result Y and the explanatory variable X, which is[1]:

$$\log(E[Y \mid X]) = \alpha + \sum_{i \in S} \beta_i \cdot x_i$$

(1)

Here \( \alpha \) represents the intercept, S is the adjustment set of the research variable, \( \beta_i \) represents the regression coefficient of the explanatory variable, and \( x_i \) represents the overall causal effect of the explanatory variable on the outcome variable Y. The final result is measured on the basis of the number of new cases per day, so the
offset $\log(A+1)$ is added to the regression model to get:

$$\log(E[Y \mid X]) = \alpha + \log(A + 1) + \sum_{i \in S} \beta_i \cdot x_i$$ (2)

The MASS package built-in function gln.nb[16], that is, the negative binomial regression fitting, gets the regression value pseudo-R2. In this experiment, it is given by $1 - \frac{V_m}{V_0}$, where $V_m$ is the sum of the squared errors of the current model, and $V_0$ is the sum of the squared error of the empty model (only intercept and offset).

6. Results

Due to the lack of part of the population mobility data, we divided the data set into two parts, and evaluated the explanatory results of model data which from January 2020 to March 15 (Table 1), and 2020 September 22 to February 24, 2020 (Table 2). It can be seen from the table that the variables have good causal explanatory properties for the epidemic (about 70%). As the epidemic has stabilized, the explanatory properties of some variables have changed (such as media index and humidity).

| Variable       | pseudo-R$^2$ | p-value             |
|----------------|-------------|--------------------|
| Mobility       | 0.717064    |                    |
| Search index   | 0.6994326   | 6.327111e-60       |
| Media index    | -0.426558   | 1.215992e-20       |
| Temperature    | 0.7052262   | 3.551134e-09       |
| Humidity       | 0.7019704   | 0.0001087921       |
| Wind           | 0.7137158   | 1.132129e-05       |

The observation period starts from the peak of the epidemic, and the prevention of the epidemic is fully conducted. Therefore, the number of COVID-19 cases reported daily continues to decline until the end of November and early December 2020, when the epidemic begins to rebound, and finally in March 2021 maintain a downward trend (Figure 3(a)). In addition, as of mid-March, the number of new COVID-19 cases per day and the total number of confirmed cases dropped sharply. Since then, the curve trend has been wavy, representing the recurrence of the new crown epidemic in various provinces and regions in China (Figure 3(b)). The light blue part of the data in the figure reflects the obvious differences in the development of the epidemic situation in
various regions.

After the sharp decline in new cases and the control of the epidemic, observational data showed that liquidity has rebounded (Figure 4(a),4(d)). At the same time, the public attention to the epidemic, especially their active vigilance, is closely related to the increase in daily cases (Figure 4(b)). In the time series throughout the year, the temperature showed a trend of rising in summer and falling in winter. At the same time, the number of cases gradually decreased with the temperature rise, and began to rebound at low temperatures (Figure 4(c)).

Within a causal framework, both the mobility and the Search index play a powerful role. In addition, temperature has also become a key factor (Figure 5).
After the sharp decline in new cases and the control of the epidemic, observational data showed that liquidity has rebounded (Figure 4(a),4(d)). At the same time, the public attention to the epidemic, especially their active vigilance, is closely related to the increase in daily cases (Figure 4(b)). In the time series throughout the year, the temperature showed a trend of rising in summer and falling in winter. At the same time, the number of cases gradually decreased with the temperature rise, and began to rebound at low temperatures (Figure 4(c)). Within a causal framework, both the mobility and the Search index play a powerful role. In addition, temperature has also become a key factor (Figure 5).

Fig. 5. Temporal variation of outcome and main determinants

References

[1] Steiger, Edgar & Mussgnug, Tobias & Kroll, Lars Eric. (2020). Causal analysis of COVID-19 observational data in German districts reveals effects of mobility, awareness, and temperature. 10.1101/2020.07.15.20154476.

[2] Schipf, S & Knüppel, Sven & Hardt, Juliane & Stang, Andreas. (2011). Directed Acyclic Graphs (DAGs) - The Application of Causal Diagrams in Epidemiology. Gesundheitswesen (Bundesverband der Ärzte des Öffentlichen Gesundheitsdienstes (Germany)). 73. 888-92. 10.1055/s-0031-1291192.

[3] Guo, Ruocheng & Cheng, Lu & Li, Jundong & Hahn, Paul & Liu, Huan. (2020). A Survey of Learning Causality with Data: Problems and Methods. ACM Computing Surveys. 53. 1-37. 10.1145/3397269.

[4] Kalisch, Markus & Zurich, Peter. (2012). Causal Inference Using Graphical Models with the R Package pcalg. JSS Journal of Statistical Software MMMMMMM YYYY. VV. 10.18637/jss.v047.i11.

[5] Textor, Johannes & Van der Zander, Benito & Gilthorpe, Mark & Liskiewicz, Maciej & Ellison, George. (2017). Robust causal inference using directed acyclic graphs: The R package ‘dagitty’. International Journal of Epidemiology. 45. 1887-1894. 10.1093/ije/dyw341

[6] Perković, Emilija & Textor, Johannes & Kalisch, Markus & Maathuis, Marloes. (2015). A Complete Generalized Adjustment Criterion. Uncertainty in Artificial Intelligence - Proceedings of the 31st Conference, UAI 2015. 10.37-48. 10.1097/00001648-199901060-00008.

[7] Tencent News. accessed 2021-02-24. URL https: https://news.qq.com/zt2020/page/feiyan.htm?#/ 

[8] Baidu Index. accessed 2021-03-03. URL https://index.baidu.com/v2/index.html#

[9] Baidu Qianxi. accessed 2021-03-22. URL https://qianxi.baidu.com/

[10] Iqbal, Najaf & Fareed, Zeeshan & Shahzad, Umer & Lina, Ma. (2020). Nexus between COVID-19, temperature and exchange rate in Wuhan City: New findings from Partial and Multiple Wavelet Coherence. Science of The Total Environment. 729. 138916. 10.1016/j.scitotenv.2020.138916.

[11] Tianqihoubao. accessed 2021-04-19. URL https: http://www.tianqihoubao.com/lishi/

[12] Wunderground. accessed 2021-04-19. URL https: https://www.wunderground.com/

[13] AQIstudy. accessed 2021-04-20. URL https: https://www.aqistudy.cn/

[14] Hongheiku. accessed 2021-03-11. URL https: https://www.hongheiku.com/

[15] National Bureau of Statistics. accessed 2020-03-11. URL https: http://www.stats.gov.cn/

[16] Greenland, Sander & Pearl, Judea & Robins, James. (1999). Causal Diagrams for Epidemiologic Research. Epidemiology (Cambridge, Mass.). 10. 37-48. 10.1097/00001648-199901000-00008.