How does COVID-19 prevalence relate to socioeconomic features?

Joshua Sia

Contents

GitHub repository link: https://github.com/joshsia/covid-socioeconomics

Audience persona

Daniel is a second-year undergraduate data science student who has mostly been taking statistics classes and has just learned about statistical modelling in R. He is familiar with regression and how to assess a model's performance, but has never worked on a data science project personally. He is interested in what data science workflows look like but not the specific code required to perform data analyses. Daniel lives in the US and the increasing number of COVID-19 cases makes him curious about the socioeconomic factors that might be associated with COVID-19 prevalence.

Abstract

COVID-19 is a serious pandemic that has introduced a wide variety of challenges since 2019. Its high transmission rate makes it difficult to control the spread of the virus and furthermore, symptoms may manifest in patients after a few days, or not manifest at all which makes it difficult for individuals to identify whether they need to self-isolate to slow down the spread of the virus. This project builds a multiple linear regression model with interaction terms to study the association of socioeconomic features with COVID-19 prevalence in the United States. By analysing the relationship between the number of COVID-19 cases and socioeconomic features of a state, factors that influence COVID-19 prevalence can be identified which could help policymakers and leaders make more informed decisions in combating COVID-19. The linear regression model with interaction terms resulted in an R-squared score of 0.67. A t-test was performed for each socioeconomic feature at the 0.1 significance level. It was found that the income ratio, the interaction between percentage of smokers and percentage of sick people, and a few other features were significantly associated with COVID-19 prevalence. Knowing aspects of a state that are associated with COVID-19 prevalence can help governments make decisions on where to allocate the state budget to attempt to reduce the number of COVID-19 cases and ultimately, to keep the state safe.

Introduction

With the recent COVID-19 pandemic, many states in the US have seen an increasing number of COVID-19 cases and associated mortailty numbers (Oster *et al.*, 2020)(team, 2020). However, controlling the spread of the virus is not simple due to its high transmission rate and furthermore, symptoms may manifest in patients after a few days, or not manifest at all which makes it difficult for individuals to identify whether they need to self-isolate to slow down the spread of the virus (Christie *et al.*, 2021)(Velavan and Meyer, 2020).

Numerous studies have been conducted on the biology of COVID-19 such as its immunopathology (Cao, 2020), pathophysiology (Yuki, Fujiogi and Koutsogiannaki, 2020) and vaccine development and efficacy (Andreadakis *et al.*, 2020). However, there are fewer papers published on the influence of socioeconomic features on COVID-19 prevalence. Previous studies have analysed the association of socioeconomics with COVID-19 prevalence across multiple countries (Chaudhry *et al.*, 2020), and the association of race, age, proportion of genders for different zip codes in the United States (Guha *et al.*, 2020). However, since there are many different ways to measure attributes of a state or country, there are still many socioeconomic factors which remain to be analysed. In addition, studying the relationship between COVID-19 prevalence and socioeconomics also has implications for fairness since the analysis may also help communities uncover underlying biases that exist in a particular area.

Thus, this project aims to analyse the relationship between COVID-19 prevalence, defined as the cumulative total number of COVID-19 cases per 100,000 people, and the percentage of smokers, income ratio, percentage of sick people, percentage of unemployed people, and the teen birth rates in the US using a simple and interpretable model.

Methods

Data

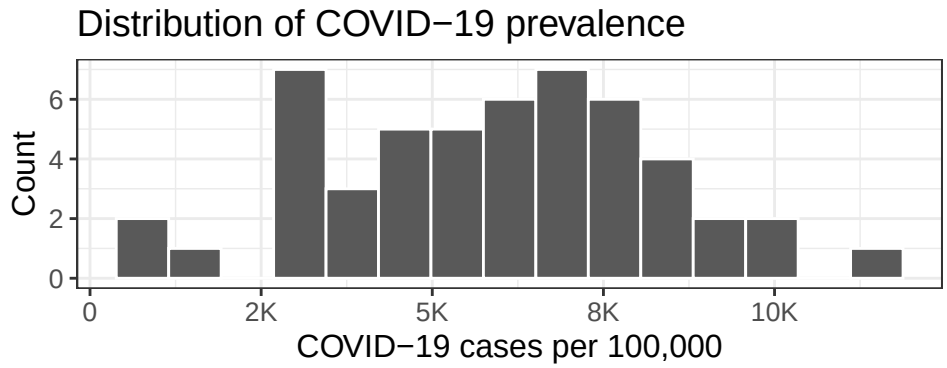
The original data set used in this project is of US socioeconomic features by county created by Dr. John Davis at Indiana University, the United States (Davis, 2020). Each row in the original data set corresponds to a date for each US county along with the new number of COVID-19 cases and socioeconomic features of the particular county. There are over 790,000 rows and over 200 features in the original data set. An arbitrary subset of features was chosen and wildcard features such as the teen birth rate and percentage of unemployed people were also selected which might be related to broader social determinants of public health.

The data set reports time series data per county for new number of COVID-19 cases and different socioeconomic features. However, due to difficulties in measurements and reporting, COVID-19 cases and socioeconomic features were updated at irregular intervals (e.g. COVID-19 cases were reported daily, whereas the socioeconomic features were reported no more than once a month). Thus, the time series data was condensed into summary statistics, specifically, the mean was calculated for socioeconomic features for each county, and the number of COVID-19 cases were summed. All missing values were removed during data wrangling since they did not account for a large proportion of the data set ($< 2\%$). The processed data set contains 51 rows corresponding to each US state, 1 column for COVID-19 prevalence, and 5 columns for different socioeconomic features.

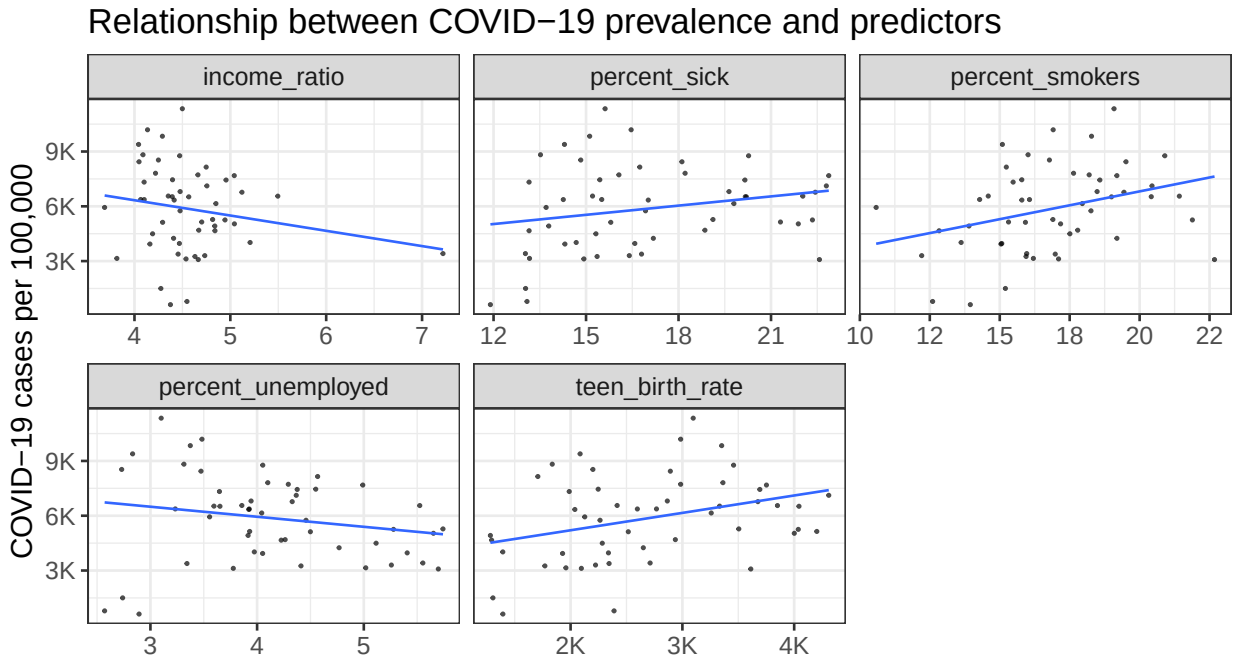
The data set used does not involve data about individuals, hence it is not sensitive, and can be found publicly on Kaggle. This data set is also used only for educational purposes in this project.

EDA

First, the empirical distribution of the number of COVID-19 cases per 100,000 was plotted as a histogram. There appears to be some observations with a very low proportion of COVID-19 cases, however, the number of COVID-19 cases per 100,000 seems to follow a normal distribution.



Next, the relationship between COVID-19 prevalence and socioeconomic features were shown as a scatterplot. A linear regression line is also shown in blue to give early hints about their association. The linear relationships do not appear strong individually, however, this could be because each feature is observed in isolation. There might be interactions between these features which can have a linear relationship with COVID-19 prevalence.



Model fitting

The response, COVID-19 cases per 100,000, appears to follow a normal distribution and there seems to be a linear relationship between the response and the explanatory variables. Thus, a multiple

linear regression (MLR) model is selected. Another advantage of the MLR is that it is simple, easily interpretable and fast. The baseline model is an MLR with only additive terms and the full model is an MLR with interaction terms. Interaction terms were included in the model to account for non-linear relationships between COVID-19 prevalence and socioeconomic features. Both models are fitted using the `lm` function from the `stats` package in `R` and compared below. Code used to perform the analysis and create this report can be found here.

In order to make interpretation of regression coefficients easier, scaling of the features was also performed such that features have a mean of zero, and a standard deviation of 1.

Results

Model selection

Model selection tools such as the adjusted R^2 value, the AIC and BIC are looked at for both models. The adjusted R^2 and the AIC favour the full model with interaction terms while the BIC favours the baseline model slightly more. This is likely because the BIC penalises an increase in model complexity more heavily than the AIC.

Model	Adjusted R squared	AIC	BIC
Baseline	0.20	933.32	946.84
Full	0.52	914.11	946.95

An F-test was carried out to determine whether the full model with interaction terms fits the data significantly better than the baseline model using the `anova` function from the `stats` package in `R`. The p-value associated with the F-statistic was 9.6×10^{-4} which is smaller than $\alpha = 0.05$. Thus, the full model with interaction terms is selected.

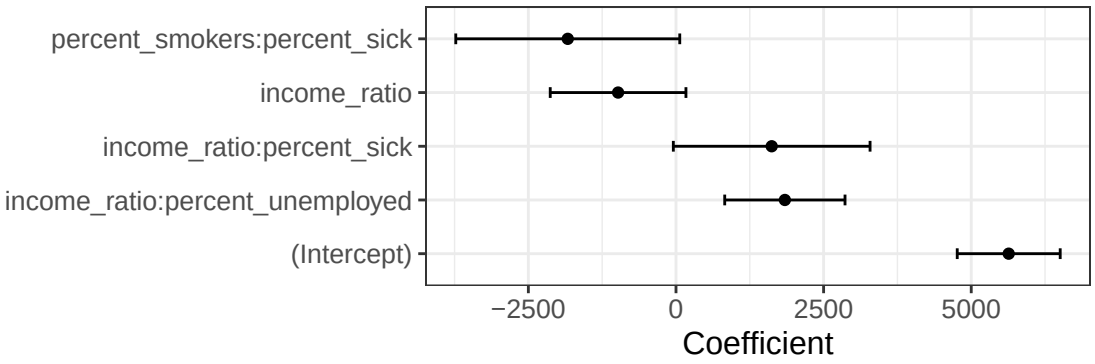
The full model with interaction terms had an R^2 score of 0.665 which implies that 66.5% of the variance in the data was explained by the model. This is a moderate score which suggests that the features used in the model may not be the best at explaining the data, or an MLR model may not be the best to model the data.

Model coefficients

Out of 16 regression coefficients in the full model, only 5 coefficients are significant at the $\alpha = 0.1$ significance level. Interestingly, most of the significant coefficients are interaction terms.

Term	Estimate	p-value
(Intercept)	5630	0.000000
income_ratio:percent_unemployed	1840	0.000791
income_ratio:percent_sick	1620	0.056200
percent_smokers:percent_sick	-1830	0.057800
income_ratio	-980	0.092100

The significant coefficients along with their 95% confidence intervals are plotted as error bars.



Conclusion

Using a multiple linear regression model with interaction terms, the socioeconomic features found to be significantly associated with COVID-19 prevalence are the income ratio, the interaction between income ratio and percentage of unemployed people, the interaction between income ratio and percentage of sick people, and the interaction between percentage of smokers and sick people.

Interestingly, the interaction between percentage of smokers and sick people is negatively associated with COVID-19 prevalence. One possible reason is that smokers are not more susceptible to contracting COVID-19 than non-smokers or smokers are more cautious about COVID-19 and tend to stay at home more than non-smokers since COVID-19 is a pulmonary virus which affects the lungs.

Furthermore, the income ratio is also negatively associated with COVID-19 prevalence, however, there is a high leverage point in the data. It is worth exploring what state this point corresponds to, and to check whether there has been a mistake in the data collection process.

It is important to note that the results of the full model should be taken with caution since the model resulted in a moderate R^2 score of 0.67. This suggests that the features chosen in this project are not the most informative at explaining COVID-19 prevalence or an MLR may not be best to model the data. Non-linear models can also be fitted and compared to the MLR models to see whether they fit the data better.

The original data set involved more than 200 features and only 5 were selected for analysis in this project due to time constraints. In the future, it would be interesting to explore how other socioeconomic features are associated with COVID-19 prevalence.

References

Andreadakis, Z. *et al.* (2020) 'The COVID-19 vaccine development landscape', *Nature reviews. Drug discovery*, 19(5), pp. 305–306.
Cao, X. (2020) 'COVID-19: Immunopathology and its implications for therapy', *Nature reviews immunology*, 20(5), pp. 269–270.
Chaudhry, R. *et al.* (2020) 'A country level analysis measuring the impact of government actions, country preparedness and socioeconomic factors on COVID-19 mortality and related health outcomes', *EClinicalMedicine*, 25, p. 100464.
Christie, A. *et al.* (2021) 'Guidance for implementing COVID-19 prevention strategies in the context of varying community transmission levels and vaccination coverage', *Morbidity and Mortality*

Weekly Report, 70(30), p. 1044.
Davis, J. (2020) 'US social determinants of health by county', *Kaggle*. Available at: https://www.kaggle.com/johnjdavisiv/us-counties-covid19-weather-sociohealth-data.
Guha, A. *et al.* (2020) 'Community and socioeconomic factors associated with COVID-19 in the united states: Zip code level cross sectional analysis', *medRxiv* [Preprint].
Oster, A.M. *et al.* (2020) 'Trends in number and distribution of COVID-19 hotspot counties—united states, march 8–july 15, 2020', *Morbidity and Mortality Weekly Report*, 69(33), p. 1127.
team, I.C. forecasting (2020) 'Modeling COVID-19 scenarios for the united states', *Nature medicine* [Preprint].
Velavan, T.P. and Meyer, C.G. (2020) 'The COVID-19 epidemic', *Tropical medicine & international health*, 25(3), p. 278.
Yuki, K., Fujiogi, M. and Koutsogiannaki, S. (2020) 'COVID-19 pathophysiology: A review', *Clinical immunology*, 215, p. 108427.