With continuously changing abilities and accumulating mutations, SARS-CoV-2, the virus that causes COVID-19, have constant evolvements and accumulated mutations in its genetic code over time. The emergence and quick spread of the alpha, beta, and delta SARS-CoV-2 VOCs have generated continuous waves of infection in the past two years. The virus has brought tremendous shocks to the supply side of the economy and resulted in millions of deaths around the globe, representing an unprecedented tragic loss of the whole human society.

By analyzing the Covid-19 Case Surveillance Public Use Data from the Centers for Disease Control and Prevention, our project aims to identify the primary factors that are sensible to the effects of Covid-19. We mainly focus on samples in North Carolina and ignore the individual observations which have missing/unknown live status records. Focusing on a single state would help eliminate potential time-invariant effects among different states.

We hope this project will bring suggestive policy implications by identifying the most vulnerable groups against the Covid-19 virus among the population. Hopefully with our convincing results, the medical facilities would be able to allocate resources, such as hospitalization and medical aids, to the appropriate groups efficiently. Also, the government can assign social welfare benefits and designate priorities for vaccination by understanding which group is most vulnerable to the virus.

Q1: What are the primary factors you intend to test initially (what treatments)?

As mentioned in the introduction, we aim to identify the most vulnerable group against Covid-19 virus. Our primary factors include characteristic variables (age, sex, race), Geological variables (County), sociological variables (hospitalization).

Q2: Any specification on the primary variables?

Age\_Group: This is a categorical variable which has three values: 0-17 years; 18-49 years; 50-64 years; 65+years

Race: This is another categorical variable which has six values: American Indian/Alaska Native; Asian; Black; Multiple/Other; Native Hawaiian/Other Pacific Islander; White

Sex: This is a variable which has three values: Male; Female; Other

Ethnicity: This is a variable which has two values: Hispanic/Non-Hispanic

Hospitalization: Was this patient hospitalized? Yes/No/Unknown

Death\_yn: Did the patient die as a result of this illness? Yes/No

Underlying\_conditions\_yn: Did the patient have one or more of the underlying medical conditions and risk behaviours: diabetes mellitus, hypertension, severe obesity (BMI>40), cardiovascular disease, chronic lung disease, other chronic diseases, immunosuppressive condition, autommune condition, current smoker, former smoker, substance abuse or misuse, disability, psychological/psychiatric, pregnancy, other.

Q3: Policy implications based on your results?

Q4: What methods, such as the linear regression model, do you implement for testing?

We use the Double-robust estimator model.

Q4: Why choose this model? What are the (dis-)advantages of choosing linear regress/decision tree/…/ over others?

Q5: How do you compare the results from different models? Which are the key coefficients/parameters you look at?

Q6: Why choose North Carolina? Why not choose the United States?

We choose North Carolina to eliminate potential bias existing in the large volume of data. When including every state in the country, the data contains 1.8 billion observations during our time interval. To reduce the time-invariant effect lying in each state, such as the geography, population structure differences and government efficiencies, which our data has no measures of, we restrict our observations to a single state: North Carolina.

Q7: Why 2020?

We focus on the year 2020 to eliminate the potential effects of vaccination. According to the reports from CDC, North Carolina's COVID-19 Vaccine eligibility opens for all adults on April 7. Since our data source does not contain information regarding the individual's vaccination status, we only focus on the year 2020.

Q8: Any limitations?

Q9: Any past literature? What are the differences between this project and them?

In the past two years, several papers have discussed the potential determinants of Covid-19 death rates. Lan Feinhandler and four other authors offer several predictors that lead to the death rate during the first eight months of 2020. They implement the OLS model/Two-stage regression model/Lasso regression model and conclude that the national Covid-19 death rate is greater than that of other flu pandemics. Also, the increase in the reported death rate in states with Democratic governors is higher than the increase in states with Republican governors. (这里记得cite). Besides, in the paper *Determinants of COVID-19 Death Rate in Europe: Empirical Analysis*, six authors use the OLS models to test multiple hypotheses. They finally prove that the population density in European countries does not affect the COVID-19 death rate. Also, the COVID-19 death rate will not drastically raise mortality statistics since people already at risk are susceptible to the disease. (这里记得cite一下)

The difference between past literature and our project is that first, we focus on a single state, and our data source covers the whole 2020 year. Besides, as mentioned above, our project uses the Double-robust estimator model.

Q10: What is your datasource?

As mentioned in the introduction part, our project uses data from CDC, Centers for Disease Control and Prevention. This case surveillance public use dataset has 19 elements for all COVID-19 cases shared with CDC. It includes demographics, geography (county and state of residence), any exposure history, disease severity indicators and outcomes, and the presence of any underlying medical conditions and risk behaviours. As for a specification of the variables, one can check the answer of Q2.

就我们可以用这个tau dr去estimatecasual effect

Reference

[1] Paper 1: **Predictors of Death Rate during the COVID-19 Pandemic**

Can check the potential factors affecting the death rate. OLS/Two-stage/Lasso

Link: <https://www.mdpi.com/2227-9032/8/3/339/htm>

[2] Paper 2: **Determinants of COVID-19 Death Rate in Europe: Empirical Analysis**

Three hypothesis, mainly linear regression

[3] Paper 3: **WHAT DOES AND DOES NOT CORRELATE WITH COVID-19 DEATH RATES**