Supplementary Material for "A Comprehensive Analysis of COVID-19 Transmission and Fatality Rates at the County level in the United States considering Socio-Demographics, Health Indicators, Mobility Trends and Health Care Infrastructure Attributes"

Independent Variable Description

For examining the critical factors contributing to the COVID-19 transmission and mortality, we compiled information on a host of exogenous variables including a) socio-demographics, (b) health indicators, (c) mobility trends, and (d) health care infrastructure. The socio-demographic variables are collected from the American Community Survey (ACS) which includes information on the age, gender, race, income, location (urban or rural), education status, income inequality and employment. With respect to health indicators, we identify the percentage of population suffering from cancer, cardiovascular disease, hepatitis, Chronic Obstructive Pulmonary Disease (COPD); diabetes, obesity, Human Immunodeficiency Virus (HIV), heart disease, kidney disease, asthma; drinking and smoking habits. The health indicator variables are collected from the Centers for Disease Control and Prevention (CDC) systems. Further, within the mobility trends, we considered two variables: daily average exposure and social distancing metric (from SafeGraph) to serve as surrogate measures for the mobility patterns. The exposure variables provide information compiled based on smartphone movement data within and across the counties in US[20]. For our analysis, we confined our attention to the overlapping movements within the counties. From the movement data provided by PlaceIQ, for each smartphone device visiting a location, the total number of distinct devices visiting that location at that particular time is calculated[20]. These distinct devices will serve as exposure for the particular device. Similarly, one can compute the exposure for all the devices residing in a county and finally compute the daily average exposure at the count level. The second measure, information on social distancing is collected from Safegraph data. These metrics provide information on the number of devices completely staying at home on a daily basis for each county. Finally, within the healthcare infrastructure attributes, we consider the hospital per capita, ICU beds per capita and COVID-19 testing rate. Information about the hospitals and ICU beds are gathered from the County level health ranking data. On the other hand, the COVID-19 testing measures are sourced from the COVID-19 tracking[29] project which provides a complete picture of the testing level as well the number of positive and negative cases for each county in the United States.

Linear Mixed Model

The two dependent variables: (a) COVID-19 daily transmission rate and (b) COVID-19 mortality rate are continuous in nature and linear regression model is the most traditional method to study such continuous responses. For the transmission rate model, we adopt the linear mixed model as there are repetitions (101 days) for each county. A detail description of the linear mixed model is provided below:

Let $q = 1, 2, \ldots, Q$ be an index to represent each county, and $d = 1, 2, \ldots, D$ be an index to represent the various days on which data (cases) was collected. The general form of the mixed linear regression model has the following structure:

$$y_{qd} = \beta X + \varepsilon_{qd}$$

where y_{qd} is the dependent variable representing the new COVID 19 cases per 100K population, X is the vector of attributes and β is the model coefficients. ε_{qd} is the random error term
assumed to be normally distributed across the dataset.

This ε term captures the dependencies across the repetition for each county. In our analysis, we estimate the correlation for different level of repetition measures: correlation for all records (101 repetitions), monthly level (31 repetitions) and weekly level (7 repetitions). The flexibility offered by the mixed model for testing dependencies enhances the model development exercise over its simpler form. In this structure, the data can be visualized as K (K = 101 or 31 or 7) records for each 1,258 counties. Estimating a full covariance matrix (up to 101*101) is computationally intensive while providing very little intuition. Hence, we parameterize the covariance matrix (Ω). For estimating a parsimonious specification, we assume a first-order autoregressive moving average correlation structure with three parameters σ, ρ, and φ as follows:

$$\Omega = \sigma^2 \begin{pmatrix} 1 & \varphi\rho & \varphi\rho^2 & \cdots & \varphi\rho^K \\ \varphi\rho & 1 & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ \varphi\rho^K & \cdots & \cdots & \cdots & 1 \end{pmatrix}$$

where, σ represents the error variance of ε, φ represents the common correlation factor across time periods K, ρ represents the dampening parameter that reduces the correlation with time and K represents the level of repetition. The correlation parameters φ and ρ, if significant, highlight the impact of county effects on the dependent variables.