Drivers of COVID-19 waves in U.S. counties, 2020–2022

Christopher F Baum (Boston College, DIW Berlin & CESIS)
Andrés García-Suaza (Universidad del Rosario, Bogotá)
Jesús Otero (Universidad del Rosario, Bogotá)
Miguel Henry (Greylock McKinnon Associates)

US Stata Conference, Washington DC, August 2022

Introduction

In this presentation, we offer a two-stage econometric modeling approach to examine a number of socioeconomic, demographic, health, epidemiological, climate and political drivers affecting the spread of COVID-19 across six pandemic waves and counties in the United States.

Our empirical strategy exploits the availability of two years of daily data: March 15, 2020 through March 19, 2022 on the number of confirmed deaths and cases of COVID-19 in 3014 U.S. counties of the 48 contiguous states and the District of Columbia. We also make use of county-level vaccination rate data for the period in which vaccinations have been available.

Introduction

In this presentation, we offer a two-stage econometric modeling approach to examine a number of socioeconomic, demographic, health, epidemiological, climate and political drivers affecting the spread of COVID-19 across six pandemic waves and counties in the United States.

Our empirical strategy exploits the availability of two years of daily data: March 15, 2020 through March 19, 2022 on the number of confirmed deaths and cases of COVID-19 in 3014 U.S. counties of the 48 contiguous states and the District of Columbia. We also make use of county-level vaccination rate data for the period in which vaccinations have been available.

In the first stage of the analysis, we use a daily-frequency panel data set on COVID-19 cases and deaths to fit mixed models of cases and deaths against lagged confirmed cases and lagged COVID-19 vaccinations for each county.

As the resulting intercept and slope coefficients are county-specific, they relax the homogeneity assumption that is implicit when the analysis is performed using geographically aggregated cross-section units.

In the first stage of the analysis, we use a daily-frequency panel data set on COVID-19 cases and deaths to fit mixed models of cases and deaths against lagged confirmed cases and lagged COVID-19 vaccinations for each county.

As the resulting intercept and slope coefficients are county-specific, they relax the homogeneity assumption that is implicit when the analysis is performed using geographically aggregated cross-section units.

In the second stage of the analysis, we assume that the county-level slope coefficient estimates are a function of factors that are taken as fixed over the course of the pandemic.

To guide the choice of regressors in the second stage, we employ the novel one-covariate-at-a-time variable selection algorithm proposed by [1].

To contrast the importance of factors over the six pandemic waves, we employ an unorthodox approach based on the seemingly unrelated regression (SUR) model.

In the second stage of the analysis, we assume that the county-level slope coefficient estimates are a function of factors that are taken as fixed over the course of the pandemic.

To guide the choice of regressors in the second stage, we employ the novel one-covariate-at-a-time variable selection algorithm proposed by [1].

To contrast the importance of factors over the six pandemic waves, we employ an unorthodox approach based on the seemingly unrelated regression (SUR) model.

In the second stage of the analysis, we assume that the county-level slope coefficient estimates are a function of factors that are taken as fixed over the course of the pandemic.

To guide the choice of regressors in the second stage, we employ the novel one-covariate-at-a-time variable selection algorithm proposed by [1].

To contrast the importance of factors over the six pandemic waves, we employ an unorthodox approach based on the seemingly unrelated regression (SUR) model.

Motivation for analysis by waves

To model the evolution of the pandemic in the U.S., we recognize that its severity has varied, given the mutating dominant variants, the introduction of widespread vaccinations, and improvements in treatment of the disease. The latter factor is particularly important as it has reduced the likelihood of mortality for those infected in most segments of the population.

A single model is not adequate to capture these variations over the past two years. We adopt the nomenclature used by the Pew Research Center in [2], who identifies six distinct waves.

Motivation for analysis by waves

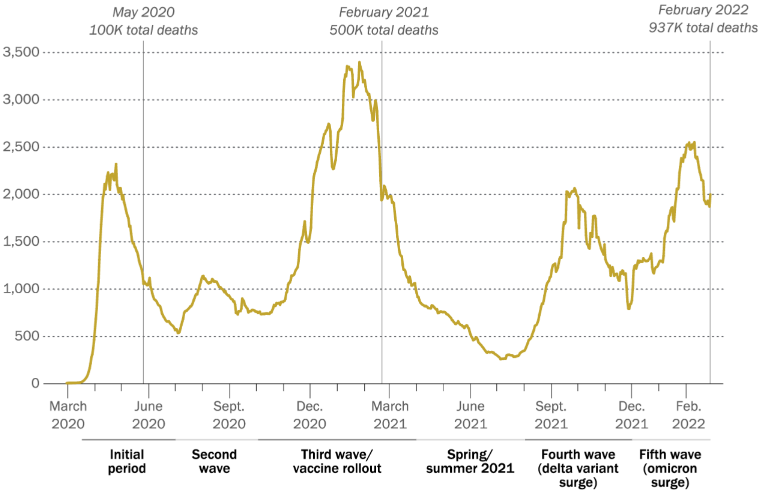
To model the evolution of the pandemic in the U.S., we recognize that its severity has varied, given the mutating dominant variants, the introduction of widespread vaccinations, and improvements in treatment of the disease. The latter factor is particularly important as it has reduced the likelihood of mortality for those infected in most segments of the population.

A single model is not adequate to capture these variations over the past two years. We adopt the nomenclature used by the Pew Research Center in [2], who identifies six distinct waves.

Motivation for analysis by waves

Two years of coronavirus deaths in the United States

Average number of daily reported coronavirus deaths in the U.S.



Notes: Seven-day rolling average number of reported COVID-19 deaths. Excludes deaths in U.S. territories and those not assigned to a specific geographic location.

Source: Pew Research Center analysis of COVID-19 data collected by The New York Times as of Feb. 28, 2022. See methodology for details.

The six waves include daily county-level data on confirmed cases and deaths for the periods:

- 1: 15 March 2020 - 30 June 2020
- 2: 1 July 2020 - 30 September 2020
- 3: 1 October 2020 - 31 March 2021
- 4: 1 April 2021 - 31 July 2021
- 5: 1 August 2021 - 30 November 2021
- 6: 1 December 2021 - 19 March 2022

Wave 3 captures the rollout of vaccines, wave 5 the surge of the delta variant, and wave 6 the dominance of the omicron variant.

The six waves include daily county-level data on confirmed cases and deaths for the periods:

- 1: 15 March 2020 - 30 June 2020
- 2: 1 July 2020 - 30 September 2020
- 3: 1 October 2020 - 31 March 2021
- 4: 1 April 2021 - 31 July 2021
- 5: 1 August 2021 - 30 November 2021
- 6: 1 December 2021 - 19 March 2022

Wave 3 captures the rollout of vaccines, wave 5 the surge of the delta variant, and wave 6 the dominance of the omicron variant.

The six waves include daily county-level data on confirmed cases and deaths for the periods:

- 1: 15 March 2020 - 30 June 2020
- 2: 1 July 2020 - 30 September 2020
- 3: 1 October 2020 - 31 March 2021
- 4: 1 April 2021 - 31 July 2021
- 5: 1 August 2021 - 30 November 2021
- 6: 1 December 2021 - 19 March 2022

Wave 3 captures the rollout of vaccines, wave 5 the surge of the delta variant, and wave 6 the dominance of the omicron variant.

The six waves include daily county-level data on confirmed cases and deaths for the periods:

- 1: 15 March 2020 - 30 June 2020
- 2: 1 July 2020 - 30 September 2020
- 3: 1 October 2020 - 31 March 2021
- 4: 1 April 2021 - 31 July 2021
- 5: 1 August 2021 - 30 November 2021
- 6: 1 December 2021 - 19 March 2022

Wave 3 captures the rollout of vaccines, wave 5 the surge of the delta variant, and wave 6 the dominance of the omicron variant.

The six waves include daily county-level data on confirmed cases and deaths for the periods:

- 1: 15 March 2020 - 30 June 2020
- 2: 1 July 2020 - 30 September 2020
- 3: 1 October 2020 - 31 March 2021
- 4: 1 April 2021 - 31 July 2021
- 5: 1 August 2021 - 30 November 2021
- 6: 1 December 2021 - 19 March 2022

Wave 3 captures the rollout of vaccines, wave 5 the surge of the delta variant, and wave 6 the dominance of the omicron variant.

The six waves include daily county-level data on confirmed cases and deaths for the periods:

- 1: 15 March 2020 - 30 June 2020
- 2: 1 July 2020 - 30 September 2020
- 3: 1 October 2020 - 31 March 2021
- 4: 1 April 2021 - 31 July 2021
- 5: 1 August 2021 - 30 November 2021
- 6: 1 December 2021 - 19 March 2022

Wave 3 captures the rollout of vaccines, wave 5 the surge of the delta variant, and wave 6 the dominance of the omicron variant.

The six waves include daily county-level data on confirmed cases and deaths for the periods:

- 1: 15 March 2020 - 30 June 2020
- 2: 1 July 2020 - 30 September 2020
- 3: 1 October 2020 - 31 March 2021
- 4: 1 April 2021 - 31 July 2021
- 5: 1 August 2021 - 30 November 2021
- 6: 1 December 2021 - 19 March 2022

Wave 3 captures the rollout of vaccines, wave 5 the surge of the delta variant, and wave 6 the dominance of the omicron variant.

The six waves include daily county-level data on confirmed cases and deaths for the periods:

- 1: 15 March 2020 - 30 June 2020
- 2: 1 July 2020 - 30 September 2020
- 3: 1 October 2020 - 31 March 2021
- 4: 1 April 2021 - 31 July 2021
- 5: 1 August 2021 - 30 November 2021
- 6: 1 December 2021 - 19 March 2022

Wave 3 captures the rollout of vaccines, wave 5 the surge of the delta variant, and wave 6 the dominance of the omicron variant.

The cumulative cases and deaths and the vaccination rates for each of the six waves are presented in the following table. We also computed these measures for two subsets of counties: those in the 4th quartile of population density, labeled High, and those in the other three quartiles, labeled Low. The impact of population density on both cases and deaths is meaningful, particularly in the earlier waves.

Table: Cumulative cases, deaths, and vaccination rates (N = 2,215,290)

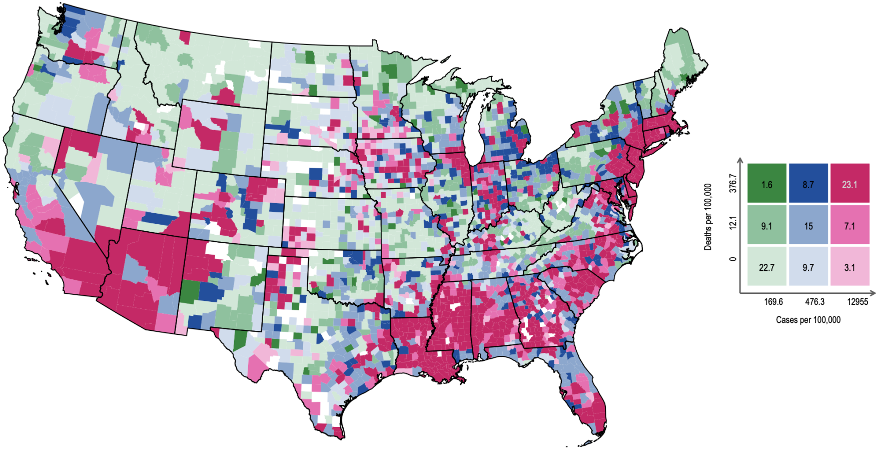
Wave starting:							
(Month/Year)		3/20	7/20	10/20	4/21	8/21	12/21
Cases/100K	Total	522.71	1962.47	9394.12	10595.04	15728.08	23992.87
	Low	479.73	1959.10	9573.60	10731.38	16117.79	24172.88
	High	651.76	1972.59	8855.23	10185.63	14557.91	23452.38
Deaths/100K	Total	17.65	43.95	189.10	211.17	284.94	357.31
	Low	13.83	41.52	199.08	222.26	303.84	379.28
	High	29.12	51.26	159.13	177.85	228.16	291.35
Vaccinations	Total	0.00	0.00	13.66	32.65	45.54	50.90
	Low	0.00	0.00	13.71	30.89	43.42	48.47
	High	0.00	0.00	13.49	37.93	51.91	58.20

To visualize the variations in the COVID-19 cumulative cases and deaths, we consider how these variables are correlated across US counties. That visualization can be implemented by Asjad Naqvi's innovative `bimap` package [3], available from the SSC Archive and documented in his Medium guide for Bi-variate maps.

We present the bivariate map of these two variables' averages for wave 1 (March–June 2020) and for wave 6 (December 2021–March 2022). The deep red color identifies the counties which are in the upper tercile of both cases and deaths, while lavender in the lower right identifies cases in the third tercile and deaths in the first tercile.

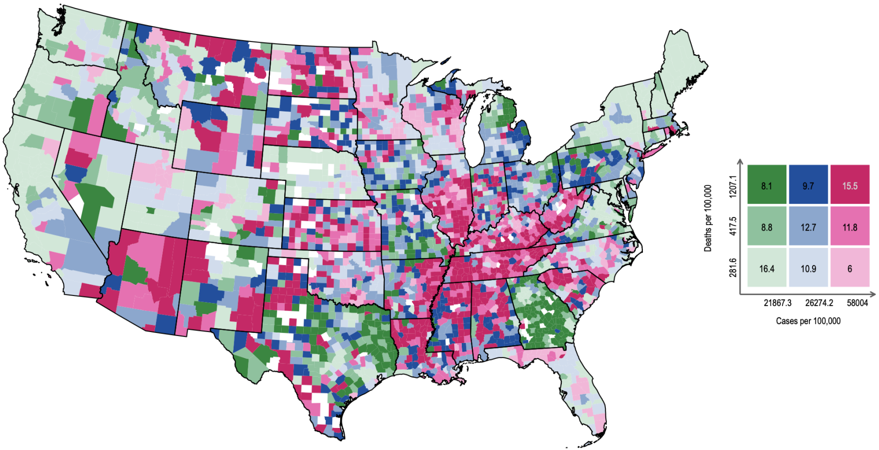
To visualize the variations in the COVID-19 cumulative cases and deaths, we consider how these variables are correlated across US counties. That visualization can be implemented by Asjad Naqvi's innovative `bimap` package [3], available from the SSC Archive and documented in his Medium guide for Bi-variate maps.

We present the bivariate map of these two variables' averages for wave 1 (March–June 2020) and for wave 6 (December 2021–March 2022). The deep red color identifies the counties which are in the upper tercile of both cases and deaths, while lavender in the lower right identifies cases in the third tercile and deaths in the first tercile.



Colors defined by tercile cutoffs of deaths and cases in counties, percentages of grand total displayed

Wave 6: December 2021–March 2022



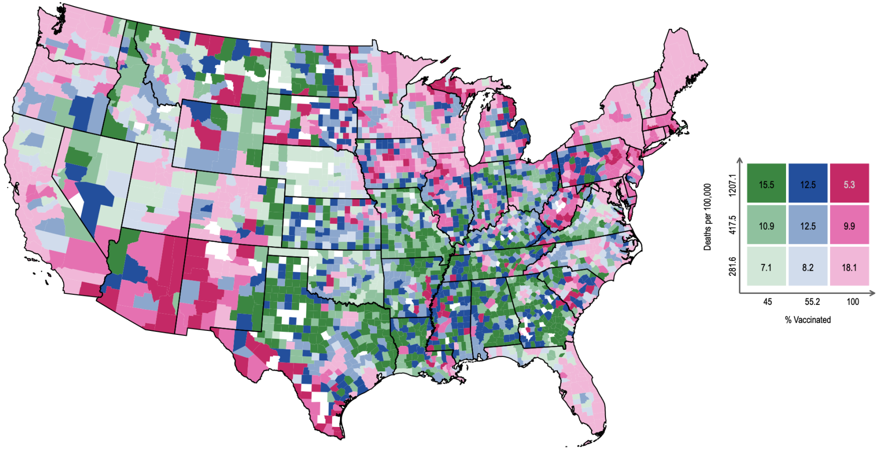
Colors defined by tercile cutoffs of deaths and cases in counties, percentages of grand total displayed

The following map illustrates the bivariate relationship between vaccination rates and cumulative deaths in wave 6.

In this map's legend, the rightmost categories refer to the third tercile of vaccination rates. The dark green counties are those with low vaccination rates and the third tercile of cumulative deaths in wave 6. The geography of those most seriously affected by the course of the virus is quite evident.

The following map illustrates the bivariate relationship between vaccination rates and cumulative deaths in wave 6.

In this map's legend, the rightmost categories refer to the third tercile of vaccination rates. The dark green counties are those with low vaccination rates and the third tercile of cumulative deaths in wave 6. The geography of those most seriously affected by the course of the virus is quite evident.



Colors defined by tercile cutoffs of deaths and vaccinated in counties, percentages of grand total displayed

First stage modeling

We now present the detailed econometric strategy for our investigation of these relationships across time and space.

We analyze the associations between cumulative confirmed cases and deaths at the county level, each expressed per 100,000 population. Starting with wave 3, we also consider the fraction of the county population recorded as being fully vaccinated.

The models are fit separately for each of the six waves to allow for variations over those episodes in the transmissibility of the virus, the impact of vaccinations, and improvements in treatment regimes.

First stage modeling

We now present the detailed econometric strategy for our investigation of these relationships across time and space.

We analyze the associations between cumulative confirmed cases and deaths at the county level, each expressed per 100,000 population. Starting with wave 3, we also consider the fraction of the county population recorded as being fully vaccinated.

The models are fit separately for each of the six waves to allow for variations over those episodes in the transmissibility of the virus, the impact of vaccinations, and improvements in treatment regimes.

In order to allow for heterogeneity within a wave, we fit mixed models (Stata's `mixed`), allowing both intercept and slopes to vary by county in this panel data context. An unstructured covariance matrix is used to provide flexibility for the random effects at the county level.

The first model for the daily county-level confirmed cases is autoregressive, with a single regressor: the county-level confirmed cases 14 days prior, capturing transmissibility of the disease. In waves 3–6, the county-level vaccination rate 14 days prior is also included.

In order to allow for heterogeneity within a wave, we fit mixed models (Stata's `mixed`), allowing both intercept and slopes to vary by county in this panel data context. An unstructured covariance matrix is used to provide flexibility for the random effects at the county level.

The first model for the daily county-level confirmed cases is autoregressive, with a single regressor: the county-level confirmed cases 14 days prior, capturing transmissibility of the disease. In waves 3–6, the county-level vaccination rate 14 days prior is also included.

The confirmed case model:

$$c_{it} = \alpha_0 + \alpha_i + \beta_0 c_{i,t-j} + \beta_i c_{i,t-j} + \gamma_0 v_{i,t-j} + \gamma_i v_{i,t-j} + \epsilon_{it}, \tag{1}$$

c_{it} and $c_{i,t-j}$ denote the cumulative confirmed cases per 100,000 in county i at time t and $t-j$, respectively. In turn, $v_{i,t-j}$ indicates the percentage of county residents that are fully vaccinated (with a second dose of a two-dose vaccine or a dose of a single-dose vaccine); α_0, β_0 and γ_0 denote unknown fixed parameters; α_i, β_i and γ_i denote county-level random effects; and ϵ_{it} is the disturbance term.

The second model associates the daily county-level deaths with a single regressor: the county-level confirmed cases 14 days prior, capturing the mortality risk for those infected. In waves 3–6, the county-level vaccination rate 14 days prior is also included.

$$d_{it} = \delta_0 + \delta_i + \kappa_0 c_{i,t-j} + \kappa_i c_{i,t-j} + \lambda_0 v_{i,t-j} + \lambda_i v_{i,t-j} + \varepsilon_{it}. \tag{2}$$

δ_0, κ_0 and λ_0 are unknown fixed parameters; δ_i, κ_i and λ_i denote county-level random effects; and ε_{it} is the disturbance term.

Although one might argue that the vaccination rate might not directly affect the likelihood that a diagnosed individual succumbs to the disease, it might be that locales with higher vaccination rates are more likely to have access to better health care. In the most recent phase of the pandemic, deaths have been concentrated among the elderly, who may have been less likely to receive vaccinations.

As a robustness test, we reestimate all first stage models using a 21-day lag of daily cases and the vaccination rate. The results are quite similar from a qualitative standpoint.

Although one might argue that the vaccination rate might not directly affect the likelihood that a diagnosed individual succumbs to the disease, it might be that locales with higher vaccination rates are more likely to have access to better health care. In the most recent phase of the pandemic, deaths have been concentrated among the elderly, who may have been less likely to receive vaccinations.

As a robustness test, we reestimate all first stage models using a 21-day lag of daily cases and the vaccination rate. The results are quite similar from a qualitative standpoint.

Following the estimation of the first stage models, the county-level random slopes for the lagged case regressor are predicted and added to the fixed coefficient for that variable.

Although the mean of county-level random effects is zero over the entire sample, it varies considerably at the county level for each wave, reflecting the heterogeneity in these dynamic relationships that arises from state-level and county-level characteristics and policies.

Following the estimation of the first stage models, the county-level random slopes for the lagged case regressor are predicted and added to the fixed coefficient for that variable.

Although the mean of county-level random effects is zero over the entire sample, it varies considerably at the county level for each wave, reflecting the heterogeneity in these dynamic relationships that arises from state-level and county-level characteristics and policies.

Second stage modeling

In the second stage, the outcome variables are the cross-sectional coefficients computed for each county and wave. An extensive set of fixed factors are considered as possible drivers of the transmissibility coefficients (from the case equation) and mortality risk coefficients (from the death equation).

The socioeconomic and demographic factors include:

- Age and sex distribution
- Racial/ethnic distribution
- Log median household income
- High school, college completion
- Poverty rate
- Rate of owner-occupied housing

The socioeconomic and demographic factors include:

- Age and sex distribution
- Racial/ethnic distribution
- Log median household income
- High school, college completion
- Poverty rate
- Rate of owner-occupied housing

The socioeconomic and demographic factors include:

- Age and sex distribution
- Racial/ethnic distribution
- Log median household income
- High school, college completion
- Poverty rate
- Rate of owner-occupied housing

The socioeconomic and demographic factors include:

- Age and sex distribution
- Racial/ethnic distribution
- Log median household income
- High school, college completion
- Poverty rate
- Rate of owner-occupied housing

The socioeconomic and demographic factors include:

- Age and sex distribution
- Racial/ethnic distribution
- Log median household income
- High school, college completion
- Poverty rate
- Rate of owner-occupied housing

The socioeconomic and demographic factors include:

- Age and sex distribution
- Racial/ethnic distribution
- Log median household income
- High school, college completion
- Poverty rate
- Rate of owner-occupied housing

The health factors include:

- Health insurance coverage
- Diabetes prevalence
- Smoking prevalence
- Years of life expectancy
- Index of severe COVID-19 health risk (z-score)
- Access to exercise opportunities
- ICU beds per capita
- Medicaid expansion status

The health factors include:

- Health insurance coverage
- Diabetes prevalence
- Smoking prevalence
- Years of life expectancy
- Index of severe COVID-19 health risk (z-score)
- Access to exercise opportunities
- ICU beds per capita
- Medicaid expansion status

The health factors include:

- Health insurance coverage
- Diabetes prevalence
- Smoking prevalence
- Years of life expectancy
- Index of severe COVID-19 health risk (z-score)
- Access to exercise opportunities
- ICU beds per capita
- Medicaid expansion status

The health factors include:

- Health insurance coverage
- Diabetes prevalence
- Smoking prevalence
- Years of life expectancy
- Index of severe COVID-19 health risk (z-score)
- Access to exercise opportunities
- ICU beds per capita
- Medicaid expansion status

The health factors include:

- Health insurance coverage
- Diabetes prevalence
- Smoking prevalence
- Years of life expectancy
- Index of severe COVID-19 health risk (z-score)
- Access to exercise opportunities
- ICU beds per capita
- Medicaid expansion status

The health factors include:

- Health insurance coverage
- Diabetes prevalence
- Smoking prevalence
- Years of life expectancy
- Index of severe COVID-19 health risk (z-score)
- Access to exercise opportunities
- ICU beds per capita
- Medicaid expansion status

The health factors include:

- Health insurance coverage
- Diabetes prevalence
- Smoking prevalence
- Years of life expectancy
- Index of severe COVID-19 health risk (z-score)
- Access to exercise opportunities
- ICU beds per capita
- Medicaid expansion status

The health factors include:

- Health insurance coverage
- Diabetes prevalence
- Smoking prevalence
- Years of life expectancy
- Index of severe COVID-19 health risk (z-score)
- Access to exercise opportunities
- ICU beds per capita
- Medicaid expansion status

The other factors include:

- Average seasonal temperatures in summer and winter
- Average seasonal relative humidity in summer and winter
- Level of $PM_{2.5}$ pollutants
- Log population density
- 2020 presidential election Democratic vote share
- Social Vulnerability Index

The other factors include:

- Average seasonal temperatures in summer and winter
- Average seasonal relative humidity in summer and winter
- Level of $PM_{2.5}$ pollutants
- Log population density
- 2020 presidential election Democratic vote share
- Social Vulnerability Index

The other factors include:

- Average seasonal temperatures in summer and winter
- Average seasonal relative humidity in summer and winter
- Level of $PM_{2.5}$ pollutants
- Log population density
- 2020 presidential election Democratic vote share
- Social Vulnerability Index

The other factors include:

- Average seasonal temperatures in summer and winter
- Average seasonal relative humidity in summer and winter
- Level of $PM_{2.5}$ pollutants
- Log population density
- 2020 presidential election Democratic vote share
- Social Vulnerability Index

The other factors include:

- Average seasonal temperatures in summer and winter
- Average seasonal relative humidity in summer and winter
- Level of $PM_{2.5}$ pollutants
- Log population density
- 2020 presidential election Democratic vote share
- Social Vulnerability Index

The other factors include:

- Average seasonal temperatures in summer and winter
- Average seasonal relative humidity in summer and winter
- Level of $PM_{2.5}$ pollutants
- Log population density
- 2020 presidential election Democratic vote share
- Social Vulnerability Index

A regression equation is fit separately to the coefficients generated for each of the six waves. In the initial specification, all factors are included in each equation.

These estimates are refined by applying the one-covariate-at-a-time (OCMT) variable selection algorithm proposed by [1] and implemented in a community-contributed routine, `ocmt`, [4], available from the SSC Archive.

A regression equation is fit separately to the coefficients generated for each of the six waves. In the initial specification, all factors are included in each equation.

These estimates are refined by applying the one-covariate-at-a-time (OCMT) variable selection algorithm proposed by [1] and implemented in a community-contributed routine, `ocmt`, [4], available from the SSC Archive.

OCMT serves as an alternative approach to penalized regression for variable selection in high-dimensional linear regression models. Its objective is to find a set of predictors that is sufficient to approximate the true data generating process underlying the variable of interest. Among the several advantages of OCMT over penalized regression methods, its authors highlight ease of interpretation, its relation to classical statistical analysis, computational speed, and good performance in small samples.

As the name implies, OCMT tests the statistical significance of all covariates one at a time and selects those whose t-statistics are in absolute value greater than a given critical value. The critical value is computed using the critical value function $c_p(K,\theta) = \Phi^{-1}\left(1 - \frac{\mathrm{p}}{2f(K,\theta)}\right)$, where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal distribution function, $f(K,\theta) = cK^{\theta}$ for some positive constant $c = 1$ and θ, the critical value exponent; $0 < \mathrm{p} < 1$ is the nominal size of the individual test statistics; and K is the number of covariates in the regression model of interest. All of the covariates that satisfy the stated condition are selected jointly to form the initial specification of the model.

In a second stage, OCMT uses this initial specification and once again tests the statistical significance of the covariates not selected before one at a time. The procedure continues until there are no more statistically significant covariates. OCMT is fast because the number of covariates bounds the number of stages required for convergence.

As the name implies, OCMT tests the statistical significance of all covariates one at a time and selects those whose t-statistics are in absolute value greater than a given critical value. The critical value is computed using the critical value function $c_p(K,\theta) = \Phi^{-1}\left(1 - \frac{\mathrm{p}}{2f(K,\theta)}\right)$, where $\Phi^{-1}(\cdot)$ is the inverse of the standard normal distribution function, $f(K,\theta) = cK^{\theta}$ for some positive constant $c = 1$ and θ, the critical value exponent; $0 < \mathrm{p} < 1$ is the nominal size of the individual test statistics; and K is the number of covariates in the regression model of interest. All of the covariates that satisfy the stated condition are selected jointly to form the initial specification of the model.

In a second stage, OCMT uses this initial specification and once again tests the statistical significance of the covariates not selected before one at a time. The procedure continues until there are no more statistically significant covariates. OCMT is fast because the number of covariates bounds the number of stages required for convergence.

The application of OCMT to the equation for each of the six waves provides a more parsimonious specification in which only selected factors are included in the equation. The OCMT estimates for each wave produce a different set of factors for those six episodes, with varying coefficients for those factors selected for multiple waves.

These cross-section estimates can be considered as a system of six equations with differing specifications and time-varying coefficients. This system of equations is then estimated with a novel application of Zellner's seemingly unrelated regression (SUR) estimator: Stata's `sureg`. The usual context for SUR is a set of OLS equations for several units (firms, industries, countries) in a time-series context. In our application of SUR the equations correspond to different time periods, and the observations are the 3,014 U.S. counties in our analysis.

The application of OCMT to the equation for each of the six waves provides a more parsimonious specification in which only selected factors are included in the equation. The OCMT estimates for each wave produce a different set of factors for those six episodes, with varying coefficients for those factors selected for multiple waves.

These cross-section estimates can be considered as a system of six equations with differing specifications and time-varying coefficients. This system of equations is then estimated with a novel application of Zellner's seemingly unrelated regression (SUR) estimator: Stata's `sureg`. The usual context for SUR is a set of OLS equations for several units (firms, industries, countries) in a time-series context. In our application of SUR the equations correspond to different time periods, and the observations are the 3,014 U.S. counties in our analysis.

The usual rationale for SUR as a systems estimator is the degree to which each equation's error process might be contemporaneously correlated with other units' errors at each *point in time*. If those correlations are sizable, SUR can yield efficiency gains relative to single-equation estimation of each equation.

In our context, the error correlations that can be exploited are those *for each county* over the six waves of the pandemic. Those correlations should be sizable, as they reflect unobservable factors at the county level that have not been captured by the time-invariant regressors selected for each wave. The degree to which these correlations increase the precision of the estimates is evaluated by the Breusch–Pagan test for independence, computed with the `sureg` option `corr`, with the null hypothesis that the 6x6 residual correlation matrix is diagonal. Under the null, this test statistic is distributed $\chi^2(m)$, where $m = 15$, the number of subdiagonal elements in the matrix. The null is strongly rejected for both applications of the SUR technique.

The usual rationale for SUR as a systems estimator is the degree to which each equation's error process might be contemporaneously correlated with other units' errors at each *point in time*. If those correlations are sizable, SUR can yield efficiency gains relative to single-equation estimation of each equation.

In our context, the error correlations that can be exploited are those *for each county* over the six waves of the pandemic. Those correlations should be sizable, as they reflect unobservable factors at the county level that have not been captured by the time-invariant regressors selected for each wave. The degree to which these correlations increase the precision of the estimates is evaluated by the Breusch–Pagan test for independence, computed with the `sureg` option `corr`, with the null hypothesis that the 6x6 residual correlation matrix is diagonal. Under the null, this test statistic is distributed $\chi^2(m)$, where $m = 15$, the number of subdiagonal elements in the matrix. The null is strongly rejected for both applications of the SUR technique.

Table: OCMT SUR cross-section results for $\hat{\beta}_i$ in cumulative case model (eq. 1) with $j = 14$

Wave starting:	3/20	7/20	10/20	4/21	8/21	12/21
Age 1-19 yrs (%)	0.00643**			0.01688***		
Age 20-39 yrs (%)	-0.00407		-0.00610***	0.00915*		
Age 40-59 yrs (%)			-0.00092		-0.00271	-0.00382***
Age 60-79 yrs (%)	0.00201		-0.00228*	0.00307		
Black (%)	-0.00266***	-0.00105*		0.00036	-0.00287***	-0.00085***
Hispanic (%)	-0.00021	-0.00177**		-0.00693***	-0.00036	-0.00069***
Male (%)	0.00007		-0.00695***	-0.00838*	-0.00105	-0.00464***
Median income (log)		-0.01190	0.04332		0.09835***	0.05423***
Social vulnerability index	0.06349**	0.04213	0.04354**	0.29297***	0.11183***	
HS completion (%)	-0.19051	0.16884	0.21472***	0.80663***	0.26546***	
Some college (%)		-0.04406	-0.09222**		-0.15706***	
Poverty rate (%)	0.28348**	-0.08389	-0.08792	-0.74218***		-0.11775**
Owner-occupied housing (%)	-0.17082**		-0.02006	0.12142		
Medicaid expansion	-0.02986***	-0.04012**	-0.00963	0.22587***	-0.04535***	-0.02412***
Uninsured (%)	-0.00115	-0.00398*	-0.00119	-0.00040	-0.00172*	-0.00556***
Diabetes rate (%)		0.00268	0.00632***	-0.00545*	0.00125	0.00116*
Smoking (%)	-0.66763***		0.12001	2.17326***	0.44826***	0.92696***
Life expectancy (years)		-0.00170	-0.00039	0.02006***	-0.00094	0.00256***
Health risk index	0.00048	-0.02014*	-0.03158***	-0.02133	0.01646***	0.00230
Access to exercise (%)	-0.00032	0.13441***	-0.00072		0.04302**	0.01037
ICU (beds per 100K)	-0.00008		-0.00001	0.00055	-0.00002	0.00011
Observations	3041					

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table: OCMT SUR cross-section results for $\hat{\beta}_i$ in cumulative case model (eq. 1) with $j = 14$

Wave starting:	3/20	7/20	10/20	4/21	8/21	12/21
IPop. density (log)	0.02034***		0.02436***	0.04162***	0.02732***	0.00946***
$PM_{\mathbf{2.5}}$		-0.00888**	0.00611***	-0.02018***	0.00025	0.00175
Summer avg. temp. (C)	-0.00072	0.01394***	-0.00719***	0.01869***	-0.00383**	
Summer rel. hum. (%)	-0.00114*			-0.00318**		-0.00076***
Winter avg. temp. (C)	0.00473***	-0.01075***	-0.00101	0.03565***	-0.00215**	0.00553***
Winter rel. hum. (%)	-0.00103	0.00186		0.00582***	0.00112	0.00180***
Democratic share 2020 (%)	0.11693**		0.08423***		0.10194***	0.09820***
Constant	0.73270***	0.04286	0.49819	-3.78922***	-0.81472***	-0.34931*
Observations	3041					

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

For the cumulative confirmed case model the OCMT selection process identifies a number of regressors that are relevant for each wave: for instance, an indicator of the state's Medicaid expansion and the percent of residents uninsured. Both of those factors have negative effects on the transmissibility of the virus in each wave, but the magnitude of those effects varies considerably across waves. A number of other factors are identified as relevant in only certain waves. The log of population density has a significant positive effect in all but wave 2, the last pre-vaccination wave. The Breusch–Pagan test for the relevance of correlations among residuals from each wave rejects its null hypothesis with a p-value of 0.00.

Table: OCMT SUR cross-section results for $\hat{\kappa}_i$ in cumulative death model (eq. 2) with $j = 14$

Wave starting:	3/20	7/20	10/20	4/21	8/21	12/21
Age 1-19 yrs (%)			-0.00100***			
Age 20-39 yrs (%)		-0.00022**	-0.00121***			-0.00003
Age 40-59 yrs (%)	0.00035*		-0.00123***			
Age 60-79 yrs (%)		0.00010	-0.00109***			
Black (%)	0.00005*				0.00002	0.00001**
Male (%)	-0.00044***	-0.00021*				-0.00005*
Median income (log)		-0.00029	-0.00382***	0.00342	-0.00172	
Social vulnerability index				0.01357**	0.00179*	
HS completion (%)				0.06545***	0.00926**	
Some college (%)				-0.00792	-0.00344	
Poverty rate (%)				-0.00873	0.00058	
Owner-occupied housing (%)						0.00316***
Uninsured (%)	-0.00021**			0.00059**	0.00015***	
Diabetes rate (%)		0.00011		0.00031	0.00009	0.00003**
Smoking (%)				0.02940	-0.02284***	
Life expectancy (years)		-0.00034***		-0.00005	-0.00018**	
Health risk index		-0.00040		0.00123	0.00062**	
Access to exercise (%)	-0.00274			-0.00308		
ICU (beds per 100K)						0.00001***
Constant	0.05068***	0.05094***	0.16734***	-0.12993	0.04313***	0.00155
Observations	3041					

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

Table: OCMT SUR cross-section results for $\hat{\kappa}_i$ in cumulative death model (eq. 2) with $j = 14$

Wave starting:	3/20	7/20	10/20	4/21	8/21	12/21
Pop. density (log)	0.00149***		0.00062***			
$PM_{2.5}$	-0.00024				0.00030***	0.00008**
Summer avg. temp. (C)			-0.00030***	0.00115***	-0.00034***	
Summer rel. hum. (%)						0.00000
Winter avg. temp. (C)		0.00012***	-0.00001	-0.00036	0.00003	
Winter rel. hum. (%)	-0.00027***					
Democratic share 2020 (%)	0.00592*					
Constant	0.05068***	0.05094***	0.16734***	-0.12993	0.04313***	0.00155
Observations	3041					

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$

For the cumulative deaths model, the OCMT selection process considers far fewer factors as influencing the mortality risk from infection in waves 1–3, corresponding to the period from March 2020–March 2021. In the following two waves (April–November 2021) several additional factors are identified as having important effects, while few factors appear in wave 6 (December 2021–March 2022), perhaps reflecting the improved treatments now available.

Summary

This study analyzed two years of daily data on COVID-19 cases and deaths at the U.S. county level. The two-stage modeling approach allows for unobservable factors to affect both the estimated tramsisibility of the virus and the mortality risk for those infected, treating each of six distinct waves of the pandemic.

The cross-sectional coefficients produced in the first stage can then identify the sociodemographic, health, climate, pollution and political factors that have played important roles in these outcomes, allowing for variations in model specification and coefficients over the six waves. This flexible approach provides considerable insight to the process by which the course of the pandemic has been affected over time and space.

Summary

This study analyzed two years of daily data on COVID-19 cases and deaths at the U.S. county level. The two-stage modeling approach allows for unobservable factors to affect both the estimated tramsisibility of the virus and the mortality risk for those infected, treating each of six distinct waves of the pandemic.

The cross-sectional coefficients produced in the first stage can then identify the sociodemographic, health, climate, pollution and political factors that have played important roles in these outcomes, allowing for variations in model specification and coefficients over the six waves. This flexible approach provides considerable insight to the process by which the course of the pandemic has been affected over time and space.

References

[1] Alxander Chudik, George Kapetanios, and M Hashem Pesaran. “A one covariate at a time, multiple testing approach to variable selection in high-dimensional linear regression models”. In: *Econometrica* 86.4 (2018), pp. 1479–1512.

[2] Bradley Jones. *The Changing Political Geography of COVID-19 Over the Last Two Years*. Tech. rep. Version022-03-03. Available at https://www.pewresearch.org/politics/2022/03/03/the-changing-political-geography-of-covid-19-over-the-last-two-years/. Pew Research Center, 2022.

[3] Asjad Naqvi. *BIMAP: Stata module to produce bivariate maps*. Statistical Software Components, Boston College Department of Economics. 2022. url: https://ideas.repec.org/c/boc/bocode/s459063.html.

[4] Héctor M. Núñez and Jesús Otero. *OCMT: Stata module to perform multiple testing approach in high-dimensional linear regression*. Statistical Software Components, Boston College Department of Economics. 2020. url: https://ideas.repec.org/c/boc/bocode/s458850.html.