Using Machine Learning to Estimate the Effect of Racial Segregation on COVID-19 Mortality in the United States

Gerard Torrats-Espinosa[a,1]

[a]Department of Sociology and Data Science Institute, Columbia University, New York, NY 10027

This manuscript was compiled on December 10, 2020

This study examines the role that racial residential segregation has played in shaping the spread of the novel coronavirus disease 2019 (COVID-19) in the US as of September 30, 2020. The analysis focuses on the effects of racial residential segregation on mortality and infection rates for the overall population and on racial and ethnic mortality gaps. To account for potential confounding, I assemble a data set that includes 50 county-level factors that are potentially related to residential segregation and COVID-19 infection and mortality rates. These factors are grouped into 8 categories: demographics, density and potential for public interaction, social capital, health risk factors, capacity of the health care system, air pollution, employment in essential businesses, and political views. I use double-lasso regression, a machine learning method for model selection and inference, to select the most important controls in a statistically principled manner. Counties that are 1 SD above the racial segregation mean have experienced mortality and infection rates that are 8% and 5% higher than the mean. These differences represent an average of 4 additional deaths and 82 additional infections for each 100,000 residents in the county. The analysis of mortality gaps shows that in counties that are 1 SD above the black-white segregation mean, the black mortality rate is 8% higher than the white mortality rate. Sensitivity analyses show that an unmeasured confounder that would overturn these findings is outside the range of plausible covariates.

COVID-19 | Racial Segregation | Machine Learning

The novel coronavirus disease 2019 (COVID-19) caused by severe acute respiratory syndrome coronavirus 2 (SARS CoV-2) infection has led the world to a public health crisis of a scale not seen in a century. Since the first case in the US was documented in the state of Washington on January 20, 2020, COVID-19 has caused more than 200,000 deaths in the US and exposed stark racial disparities in disease risk and fatality rates (1). In Chicago and Milwaukee, 70% and 73% of COVID-19 deaths have been black. In the states of Louisiana and Michigan, 70% and 40% of deaths have also been black residents (2). These rates are more than twice the percentage of black individuals that make up the population in these areas.

In the United Kingdom, black individuals are 4.2 times more likely to die from COVID-19 than whites, and when adjusting fatality rates for measures of self-reported health and disability, blacks are still 1.9 times more likely to die from COVID-19 than whites (3). While death rates by racial group that adjust for pre-existing health conditions have not been reported in the US, it is plausible to think that an important part of the racial gap in mortality will remain unexplained after individual risk factors and comorbidities are taken into account.

This study examines how racial residential segregation has impacted COVID-19 mortality in a sample of 2,174 counties in the US that include approximately 96% of the country population. The link between racial segregation and health outcomes has been extensively documented in the social and medical sciences (4, 5). In their systematic review of 39 studies of segregation and health outcomes, Kramer and Hogue (4) report that, for the most part, studies that find an association between black-white segregation and health outcomes show that segregation is more detrimental for blacks than it is for whites (6–9); although in five of the studies that they review, whites are also harmed by the levels of black-white segregation in their communities (10–14). Other studies have found that the conditions of isolation and population concentration that characterize racially segregated cities have been associated with the spread of infectious diseases such as tuberculosis and HIV/AIDS among racial minorities (15, 16). Racial segregation has also been linked to higher exposure to air pollutants and environmental hazards that affect the respiratory system and increase the risk of mortality among racial minorities (17, 18).

In racially segregated counties, COVID-19 mortality among racial and ethnic minorities may be higher if racial and ethnic minorities are more exposed to contextual factors that make them more vulnerable to a highly contagious virus such as SARS-CoV-2. In addition to being more likely to experience pre-existing health conditions that increase their vulnerability to COVID-19 (19–22), blacks and Hispanics are overrepre-

Significance Statement

This study examines the role that racial residential segregation has played during the first nine months of the novel coronavirus disease 2019 (COVID-19) pandemic in the United States. To account for other factors that may explain COVID-19 mortality and infection, I assemble a data set that includes 50 county-level factors that measure demographics, density and potential for public interaction, social capital, health risk factors, capacity of the health care system, air pollution, employment in essential businesses, and political views. I use machine learning methods to guide the selection of the most important controls. Results show that more segregated counties had higher mortality and infection rates overall and larger mortality rates among blacks relative to whites.

G T-E designed research, performed research, analyzed data, and wrote the paper.

The author declares no conflict of interest.

[1]To whom correspondence should be addressed. E-mail: gerard.torrats@columbia.edu

sented in jobs that have been classified as essential during the pandemic (23), are more likely to live in multigenerational households (24), are less likely to have health insurance (25), and live in neighborhoods where essential establishments such as pharmacies and grocery stores are more scarce (26).

Highly clustered friendship networks and social interactions within members of the same racial and ethnic group may increase the risk of infection among minorities as well (27, 28). If racial minorities are at higher risk of developing COVID-19 because of the higher prevalence of underlying health conditions or because of the environment that surrounds them, more frequent within-race interactions could lead to higher mortality rates and faster transmission rates in places where minorities are more spatially concentrated.

If these mechanisms are at play, the mortality rate among disadvantaged groups like blacks and Hispanics should be higher than that of whites in places where blacks and Hispanics are more segregated from whites. However, it is also plausible to think that higher levels of segregation could translate into higher mortality rates for the overall population if the share of minorities in the county is large and thus drive the overall count of deaths and infections, or if minorities and whites overlap in public spaces (e.g., public transit and restaurants) so that the virus spills over from minority clusters to the rest of the population through these encounters.

To account for factors that may confound the relationship between residential segregation and COVID-19 mortality, I assemble a data set that includes 50 county-level factors that are potentially related to residential segregation and COVID-19 infection and mortality. These factors are grouped into 8 categories: demographics, density and potential for public interaction, social capital, health risk factors, capacity of the health care system, air pollution, employment in essential businesses, and political views. I use double-lasso regression (29, 30), a machine learning method for model selection and inference, to select the most important predictors in these categories in a statistically principled manner. The double-lasso estimates show that counties that are 1 SD above the segregation mean have experienced a mortality rate that is 8% higher, an infection rate that is 5% higher, and a gap in black-white mortality that is 8% larger. Sensitivity analyses show that an unmeasured confounder that would overturn these findings is outside the range of plausible explanations.

Racial Segregation in the COVID-19 Context

Racial residential segregation has been linked to a wide range of public health outcomes. Higher black-white segregation is associated with elevated mortality among black individuals in multiple age groups (7, 31). Deaths by homicide among blacks are higher in racially segregated metropolitan areas (32–34). Beyond mortality and life expectancy outcomes, higher racial segregation is related to higher incidence of tuberculosis and cardiovascular disease among racial minorities (15, 35), lower availability of food establishments serving healthy foods (36), and higher exposure to toxic air pollutants (17, 18).

A large literature on racial and ethnic homophily in social interactions has documented that individuals are more likely to form relationships with others of the same racial or ethnic group (27, 28, 37–40). In a highly segregated setting, these patterns of intra-group interactions may be more frequent since physical distance between groups makes inter-group connections more difficult. If racial and ethnic minorities are at higher risk of developing COVID-19, the propinquity mechanism may increase rates of infection and mortality in counties where at-risk minorities are more segregated and isolated. These more frequent interactions within minorities that are more vulnerable to COVID-19 may also increase their mortality rates relative to whites.

Racial and ethnic differences in COVID-19 infection rates have emerged in different settings. In US cities and states that report data by race and ethnicity, blacks and Hispanics are overrepresented in the death and infection counts (2, 41). A fraction of the observed racial and ethnic gaps in infection and mortality rates is likely due to underlying health conditions that make blacks and Hispanics more vulnerable to COVID-19. A study with 5,700 patients hospitalized with COVID-19 in New York City found that hypertension, obesity, and diabetes were the most common comorbidities (42). Although this study did not report the race and ethnicity of the patients, these health conditions are more prevalent among blacks and Hispanics (19–22).

Beyond pre-existing medical conditions, racial and ethnic minorities may be at higher risk of developing COVID-19 if they experience contextual factors that make them more vulnerable to the virus or make them less likely to get medical care. Blacks and Hispanics are overrepresented in jobs that have been classified as essential during the pandemic (23), are more likely to live in multigenerational households (24), are less likely to have health insurance (25), and live in neighborhoods where essential establishments such as pharmacies and grocery stores are more scarce (26). Furthermore, racial and ethnic minorities may be reluctant to seek medical care or get tested if they fear interacting with the public health system because they are undocumented or an arrest warrant has been issued on them (43, 44).

The key idea is that different individual-level likelihoods to develop COVID-19 combined with racial and ethnic homophily in social interactions may produce mortality rates that look very different depending on how the population is spatially distributed. The data collection and empirical strategy described below are designed to test these predictions.

For this study, I measure racial residential segregation using the Relative Diversity Index, a multi-group segregation metric capturing the ratio of within-tract diversity to total diversity in the county (Section 2.1 in the *SI Appendix* explains how the Relative Diversity Index is calculated from data at the Census tract level). The Relative Diversity Index can be interpreted as one minus the ratio of the probability that two individuals from the same tract are members of different racial/ethnic groups to the probability that any two individuals are members of different groups (45). The index can take values from 0 to 1, with 0 indicating that all tracts have the exact same diversity than the county as a whole (i.e., the shares of each racial group are the same across all tracts in the county), and 1 representing a county where tracts have no diversity (e.g., one set of tracts includes all black residents and no one else, another set of tracts includes all Hispanic residents and no one else, and so on). Fig. S1 in the *SI Appendix* reports bivariate associations between the 50 county attributes and the Relative Diversity Index. Table S1 in the *SI Appendix* reports the mean, standard deviation, minimum, median, and maximum values for this index. Section 3 in the *SI Appendix* reports results

using the Theil's Information Theory Index as the measure of segregation, which lead to the same conclusions than the results shown here (the Relative Diversity Index and the Theil Index have a correlation of .96 and an R^2 of .93, as shown in Fig. S2 in the *SI Appendix*).

Results

Correlates of COVID-19. To account for other county characteristics that may explain higher mortality and infection rates in racially segregated counties, I bring together data from 12 different sources to measure 50 attributes of the counties and their population. I group these 50 county characteristics into the following 8 categories: demographics, density and potential for public interaction, social capital, health risk factors, capacity of the health care system, air pollution, employment in essential businesses, and political views. The choice of these 8 categories is informed by recent evidence on what we know from the impact of COVID-19 across the US population (46), by prior studies on the social determinants of health (47, 48), and by sociological evidence on urban inequality, segregation, and health outcomes (49, 50).

Fig. 1 reports standardized bivariate associations between each of the 50 covariates and COVID-19 mortality, net of state fixed effects. Table S1 in *SI Appendix* reports the mean, standard deviation, minimum, median, and maximum values for these covariates. Table S2 in *SI Appendix* lists all data sources and explains how variables are constructed and when they are measured (all covariates are measured at some point between 2016 and 2020, based on their most recent availability). The analytic sample includes 2,174 counties for which data on all these characteristics are available. This set of counties included 96% of the US population in 2018.

For the most part, the bivariate associations in Fig. 1 reflect the higher incidence of COVID-19 in urban counties. COVID-19 mortality has been higher in counties with larger shares of racial and ethnic minorities, higher poverty rates, higher income segregation, more population density, higher share of the population living in overcrowded housing units, more traffic of passengers in nearby airports, more air pollution, and more progressive political views. These county attributes also correlate strongly with racial segregation (see Fig. S1 in the *SI Appendix*); therefore, accounting for them will be key to accurately characterize the relationship between racial segregation and COVID-19 outcomes. The empirical strategy described in the *Materials and Methods* section provides a statistically principled way to choose the most important controls. It will also enable assessing the magnitude of an unmeasured confounder that would change the estimated relationship between segregation and COVID-19 mortality rates.

Racial Segregation and Aggregated COVID-19 Mortality and Infection. Ignoring possible confounding factors, the relationship between multi-group racial segregation and COVID-19 mortality and infection is strong. Fig. 2 plots the bivariate standardized association between the Relative Diversity Index of segregation and the log rates of COVID-19 deaths and infections per 100,000 residents, net of state fixed effects. A 1 SD difference in the Relative Diversity Index of segregation predicts a 28% difference in mortality rates and a 17% difference in infection rates across the 2,174 counties in the sample.

Fig. 3 shows OLS and double-lasso estimates of the effect of segregation on COVID-19 mortality and infection as of September 30, 2020. For OLS regressions, I show estimates when no controls are included and estimates from models that include state fixed effects. The double-lasso regressions include the set of controls selected by the lasso approach described in the *Materials and Methods* section. The double-

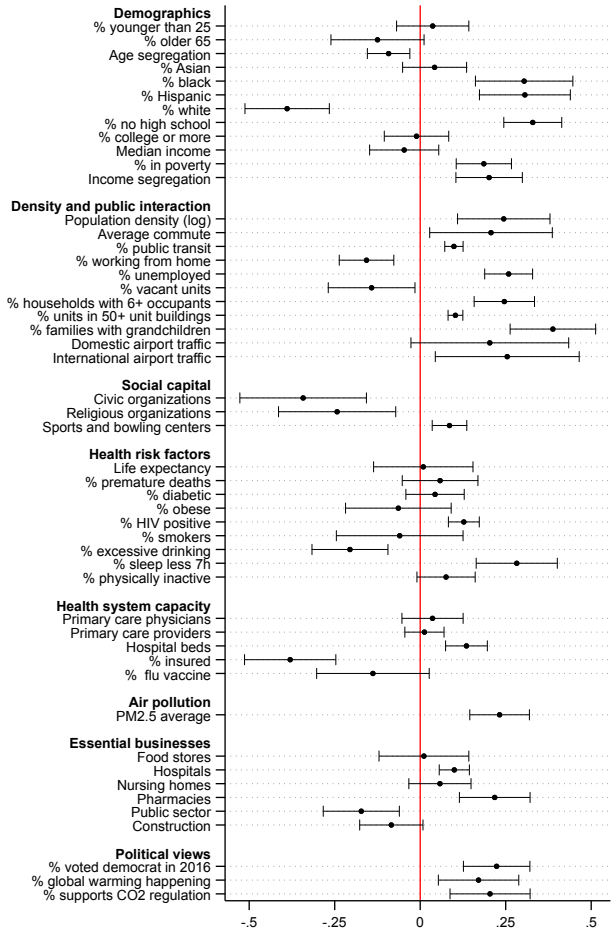

Fig. 1 Standardized bivariate associations between county attributes and COVID-19 mortality

The death rate is the log of the number of deaths per 100,000 residents in the county as of September 30, 2020. The associations are estimated via OLS regression with state fixed effects and standard errors clustered by state. Bars around estimated bivariate associations reflect 95% confidence intervals. The outcome and covariates have been standardized to have mean 0 and SD 1.

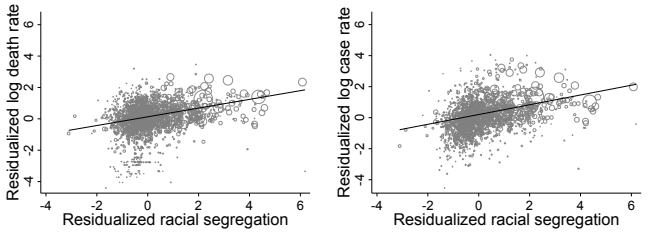

Fig. 2 Relationship between racial segregation and COVID-19 mortality and infection rates net of state fixed effects

The x-axis represents the residuals from a regression of the Relative Diversity Index of segregation on the set of state dummies. The y-axis represents the residuals from a regression of the corresponding COVID-19 outcome on the set of state dummies. The size of the dots is proportional to the county population.

lasso procedure selects 18 controls among the possible 50 candidates (in addition to the state indicators, which are forced to enter the model). Table S3 in the *SI Appendix* lists all controls and shows the full regression output for the double-lasso models.

On the basis of the double-lasso estimates in Fig. 3, a 1 SD difference in segregation predicts an overall mortality rate that is 8% higher and an overall infection rate that is 5% higher. These percentage changes represent an average of 4 additional deaths and 82 additional infections for each 100,000 residents in the county.

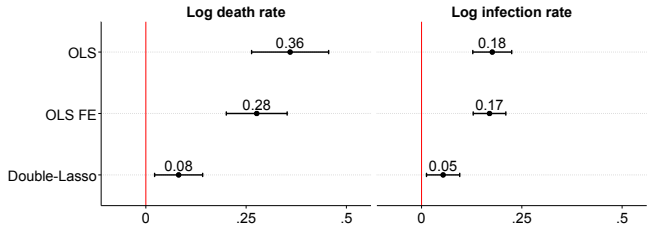

Fig. 3 OLS and double-lasso regression estimates of the effect of racial segregation on COVID-19 death and infection rates

OLS models include no controls. OLS FE models include state fixed effects. Double-lasso models include the 18 controls selected by the lasso procedure (shown in Table S3 in the *SI Appendix*) and state fixed effects. Standard errors are clustered by state. Bars around estimated coefficients reflect 95% confidence intervals. Segregation and covariates have been standardized to have mean 0 and SD 1. Outcomes are in log rates. The sample includes 2,174 counties.

Racial Segregation and Racial and Ethnic Gaps in COVID-19 Mortality. The next set of analyses focuses on the impact of the black-white Relative Diversity Index and the Hispanic-white Relative Diversity Index on black-white and Hispanic-white mortality gaps, respectively, as of September 30, 2020. Using two subsamples of counties that reported COVID-19 deaths for whites, blacks (N=243), and Hispanics (N=218), I estimate OLS and double lasso models analogous to those in Fig. 3 in which the outcomes are the difference in the death rate between blacks and whites and between Hispanics and whites. Fig. 4 shows OLS and double-lasso estimates of the effect of differences in black-white and Hispanic-white segregation on differences in black-white and Hispanic-white mortality gaps. The lasso estimates show that in counties that are 1 SD above the black-white segregation mean, the mortality rate among blacks is 8% higher than the white mortality rate. There is no association between Hispanic-white segregation and the Hispanic-white mortality gap. The full regression output from the lasso models is shown in Table S4 in the *SI Appendix*.

Robustness Checks and Sensitivity to Unmeasured Confounding. Tables S3 and S4 in the *SI Appendix* show regression results when the racial and ethnic composition of the county is fully accounted for (note that the lasso only selects % Hispanic and % white as controls) and when racial and ethnic differences in poverty rates, household median income, unemployment, and life expectancy are included in the regressions. As columns 2 and 5 show, results are unchanged when % Asian and % black are added to the set of controls. Columns 3 and 6 show that controlling for racial and ethnic differences in socioeconomic status and life expectancy does not change the conclusions of the main findings either. Figs. S4 to S6 in the *SI Appendix* are the counterparts to Figs. 2 to 4 when

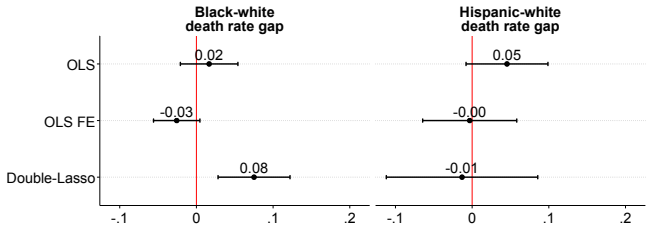

Fig. 4 OLS and double-lasso regression estimates of the effect of racial segregation on racial gaps in COVID-19 death rates

OLS models include no controls. OLS FE models include state fixed effects. Double-lasso models include the 18 controls selected by the lasso procedure (shown in Table S4 in the *SI Appendix*) and state fixed effects. Bars around estimated coefficients reflect 95% confidence intervals. Segregation and covariates have been standardized to have mean 0 and SD 1. Segregation is measured with the black-white and Hispanic-white Relative Diversity Indices. The black-white (hispanic-white) death rate gap is the difference between the log death rate for blacks (hispanics) and the log death rate for whites. The sample includes 243 counties for the black-white model and 218 for the Hispanic-white model. Figures S9 and S10 in the *SI Appendix* show results when using equal samples (N=180) across both models.

the Theil's Information Theory Index is used to measure of segregation. These models yield very similar estimates for the effect of segregation on overall mortality and infection rates and on racial/ethnic mortality gaps.

To assess the extent to which the estimate from the black-white model in Fig. 4 is biased and driven by an unobserved confounder, I implement two sensitivity tests proposed by Frank (51) and Oster (52). Fig. S7 in the *SI Appendix* shows that an omitted variable that would invalidate the inference in the black-white model (i.e., make the estimate statistically non-significant at the 5% level) would have to be more strongly correlated with the black-white mortality gap and the black-white Relative Diversity Index than any of the 50 covariates listed in Fig. 1. Similarly, Fig. S8 in the *SI Appendix* shows that for the true effect of black-white segregation on the black-white mortality gap to be zero (i.e., the point estimate is zero), there should exist an unobserved covariate that when added to the regression increases the R^2 from .56 (as shown in Table S4 in the *SI Appendix*) to .80 *and* is six times more predictive of segregation than the 18 controls selected by the lasso. These two tests point to highly implausible scenarios where the findings from Fig. 4 would be overturned.

Discussion

The history of the United States is filled with instances where natural disasters and public health crises have disproportionately upended the lives of racial and ethnic minorities and their communities (53). COVID-19 has been a continuation of this pattern. In places that report mortality rates by race and ethnicity, blacks and Hispanics are overrepresented among the infected, hospitalized, and death due to COVID-19. This study builds on the literature on the social determinants of health (47, 48) to examine how racial residential segregation has impacted mortality and infection rates across US counties. To account for potential confounders of the relationship between segregation and COVID-19 ouctomes, I have assembled a data set with 50 covariates and used a machine learning method for model selection and inference, double-lasso regression, to chose the most important ones, without making strong a-priory assumptions about the functional form.

The findings point to racial segregation as an important

driver of mortality and infection rates across the country. A 1 SD difference in racial segregation (measured with the Relative Diversity Index) predicts an overall mortality rate that is 8% higher and an overall infection rate that is 5% higher. These percentage changes represent an average of 4 additional deaths and 82 additional infections for each 100,000 residents in the county. The analyses of mortality by race and ethnicity show that in counties that are 1 SD above the black-white segregation mean, the mortality rate among blacks is 8% higher than the white mortality rate. Hispanics don't exhibit higher COVID-19 mortality rates than whites in counties where they are more segregated, a finding that is in line with extensive epidemiological evidence on the so called "Hispanic paradox" (54). Although the research design does not allow for a direct causal interpretation of these findings, sensitivity analyses reveal that an unobserved confounder that would overturn these findings is highly implausible.

The aggregated nature of the data does not allow for a test of mechanisms linking segregation and COVID-19 mortality. In the introduction, I have speculated that racial homophily in social ties may facilitate the spread of a highly contagious virus such as SARS CoV-2 if racial minorities are segregated and at higher risk of being infected. Results from models that examine mortality gaps between minorities and whites are consistent with this hypothesis, but future studies using individual-level data that tract interactions between individuals at a more granular level should extend this analysis.

Due to data limitations, the estimation of the effect of segregation on racial/ethnic gaps in mortality is not yet possible for all counties. When more data reporting deaths by race become available, extending the analyses presented here should be a priority. Looking at where COVID-19 has had its greatest impact and considering the extensive research that documents the negative effects of segregation on the health outcomes of racial minorities, it is plausible to assume that segregation will be a major contributor to racial gaps in mortality across the country.

The findings from this study should encourage epidemiologists to include features of the built environment in their mathematical models forecasting the spread of diseases. Compartmental models such as the SIR model (55) are powerful first-order approaches to predict the evolution of infectious diseases, but there is ample room to add complexity and heterogeneity to these models to better reflect the spatial clustering of population groups.

Materials and Methods

Double-Lasso Regression. Theory and prior evidence have guided us in creating a data set with possible county attributes that we think should be included in the models. But from a statistical standpoint, it is unclear whether all of them should be included in a traditional OLS regression. Covariates strongly correlated with COVID-19 outcomes and segregation are necessary to avoid omitted variable bias. Including covariates that correlate primarily with the outcome will remove residual variation in the outcome and ensure a more precise estimate of the segregation coefficient. Including covariates that strongly correlate with segregation but not so much with the outcome will unnecessarily inflate the variance of the estimated coefficient on segregation. And given the nature and novelty of the outcome, the mortality and spread of a new virus, we may not have clear priors as to which variables must be included. Leaving model selection up to the researcher could raise the suspicion that the selected controls are the ones that fit the statistically significant story that we want to tell.

Recent developments in the econometrics literature leverage the power of machine learning to guide the principled selection of variables when these are many and the correct functional form is unknown. Belloni, Chernozhukov, and Hansen (29) propose a "double-lasso" approach that identifies the relevant covariates to be included. The method uses lasso regression (56) to select covariates in two stages, first choosing those that predict the outcome and then those that predict the independent variable of interest.

Before going over the two steps in the double-lasso regression, it is important to understand the advantage that the lasso methodology has over OLS regression. Given a model with p independent variables of this form:

$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j + \epsilon, \qquad [1]$$

the OLS solution would find the vector of β coefficients that minimizes the following objective function:

$$\hat{\beta}^{OLS} = \arg\min_{\beta} \left\{ \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 \right\}. \qquad [2]$$

If X_1 is the segregation covariate, $\hat{\beta}_1$ will be the estimate of the association between segregation and the corresponding COVID-19 outcome. The problem we face is that the number of p covariates could be very large; and some of these controls may be redundant, some will be irrelevant, and for some others we may not have any priors as to whether they should be in the model.

The lasso provides us with a framework to perform a statistically principled selection of the controls that must feature in the model. Instead of minimizing the objective function in Equation (2), the lasso shrinks the regression coefficients toward zero by adding a penalty term. The penalty term, known as the $L1$-norm, forces the sum of the absolute values of the regression coefficients to be as small as possible, which reduces overfitting by discouraging complex models. Because of this constraint, the lasso will select the independent variables that contribute the most to minimizing the sum of squared errors, dropping those that contribute nothing and optimally shrinking the rest. Formally:

$$\hat{\beta}^{lasso} = \arg\min_{\beta} \left\{ \frac{1}{2} \sum_{i=1}^{N} \left(y_i - \beta_0 - \sum_{j=1}^{p} x_{ij} \beta_j \right)^2 + \lambda \sum_{j=1}^{p} |\beta_j| \right\}. \qquad [3]$$

In Equation (3), λ is the tuning parameter and controls the "strength" of the constraint that the sum of the absolute values of the regression coefficients must be as small as possible. Different values of λ will yield different sets of coefficients in β^{lasso}. If we set λ to zero, $\beta^{lasso} = \beta^{ols}$. If $\lambda = \infty$, all coefficients are set to zero. So, choosing the optimal value of λ is crucial. That choice is usually made by cross-validation over a set of subsamples of roughly equal size. Cross-validation will guard us against over-fitting the data and ensure that we maximize the predictive power of our model if we are given a new sample.

The lasso is a powerful regression tool if we are interested in predicting the outcome, but it limits us in a fundamental way if we are interested in making inferences about any of the coefficients in β^{lasso}. By allowing the lasso to shrink all coefficients toward zero (and setting some to zero), it is possible that the coefficient on segregation, β_1, will be shrunk and thus biased. Similarly, some of the coefficients on the controls that must be in the model to avoid the omitted variable bias problem could also be shrunk, which would also bias the estimate of β_1 (some of these controls could be excluded altogether if they have a small predictive power on the outcome).

The double-lasso regression is designed to solve this problem in two intuitive steps. The idea is to fit two separate lasso regressions, one of the outcome Y on all covaries and another of the independent

variable of interest, X_1, on all covariates. These two regressions will yield two sets of non-zero coefficients, one for each of the respective lassos. The final step is to fit a linear regression of the outcome Y on X_1 and the set of covariates with non-zero coefficients selected by any of the two lassos. In more concrete terms and taking Equation (1) as our model of interest, where X_1 is the measure of segregation and β_1 represents the relationship between segregation and COVID-19 mortality, the double-lasso regression proceeds as follows:

- Fit a lasso of Y on $\mathbf{X}$, where $\mathbf{X}$ includes all covariates except segregation, $X_2, ..., X_p$. Identify the set of covariates with non-zero coefficients in this regression and label it $\mathbf{X_y}$.
- Fit a lasso of X_1 on $\mathbf{X}$. Identify the set of covariates with non-zero coefficients in this regression and label it $\mathbf{X_s}$.
- Select the union of covariates in $\mathbf{X_y}$ and $\mathbf{X_s}$ and label it $\mathbf{X_u}$.
- Fit a linear regression of Y on X_1 and $\mathbf{X_u}$ to obtain an estimate of β_1.

We can intervene in the process of selecting covariates by forcing some of them to be included in $\mathbf{X_u}$. We may want to do that if we know that some covariates are absolutely necessary to obtain an unbiased estimate of β_1. In this case, I force the state indicators to be included as controls in $\mathbf{X_u}$. The covariates selected by the lasso procedure are shown in Table S3 in the *SI Appendix*.

ACKNOWLEDGMENTS. The author thanks Shamus Khan, Bruce Western, Josh Whitford, and Andreas Wimmer for their helpful comments.

1. Hooper MW, Nápoles AM, Pérez-Stable EJ (2020) COVID-19 and racial/ethnic disparities. *JAMA*.
2. Yancy CW (2020) COVID-19 and African Americans. *JAMA* 323(19):1891–1892.
3. UK ONS (2020) Coronavirus (COVID-19) related deaths by ethnic group, England and Wales. *Office for National Statistics*.
4. Kramer MR, Hogue CR (2009) Is segregation bad for your health? *Epidemiologic Reviews* 31(1):178–194.
5. Williams DR, Jackson PB (2005) Social sources of racial disparities in health. *Health Affairs* 24(2):325–334.
6. Hart KD, Kunitz SJ, Sell RR, Mukamel DB (1998) Metropolitan governance, residential segregation, and mortality among African Americans. *American Journal of Public Health* 88(3):434–438.
7. Hearst MO, Oakes JM, Johnson PJ (2008) The effect of racial residential segregation on black infant mortality. *American Journal of Epidemiology* 168(11):1247–1254.
8. Cooper RS, et al. (2016) Relationship between premature mortality and socioeconomic factors in black and white populations of us metropolitan areas. *Public Health Reports*.
9. Jackson SA, Anderson RT, Johnson NJ, Sorlie PD (2000) The relation of residential segregation to all-cause mortality: A study in black and white. *American Journal of Public Health* 90(4):615.
10. Collins CA (1999) Racism and health: segregation and causes of death amenable to medical intervention in major us cities. *Annals of the New York Academy of Sciences* 896(1):396–398.
11. Fang J, Madhavan S, Bosworth W, Alderman MH (1998) Residential segregation and mortality in new york city. *Social Science & Medicine* 47(4):469–476.
12. LeClere FB, Rogers RG, Peters KD (1997) Ethnicity and mortality in the United States: individual and community correlates. *Social Forces* 76(1):169–198.
13. Masi CM, Hawkley LC, Piotrowski ZH, Pickett KE (2007) Neighborhood economic disadvantage, violent crime, group density, and pregnancy outcomes in a diverse, urban population. *Social science & medicine* 65(12):2440–2457.
14. Rodriguez RA, et al. (2007) Geography matters: relationships among urban residential segregation, dialysis facilities, and patient outcomes. *Annals of Internal Medicine* 146(7):493–501.
15. Acevedo-Garcia D (2001) Zip code-level risk factors for tuberculosis: Neighborhood environment and residential segregation in New Jersey, 1985-1992. *American Journal of Public Health* 91(5):734.
16. Fennie KP, Lutfi K, Maddox LM, Lieb S, Trepka MJ (2015) Influence of residential segregation on survival after AIDS diagnosis among non-hispanic blacks. *Annals of Epidemiology* 25(2):113–119.
17. Downey L (2007) US metropolitan-area variation in environmental inequality outcomes. *Urban Studies* 44(5-6):953–977.
18. Morello-Frosch R, Jesdale BM (2006) Separate and unequal: Residential segregation and estimated cancer risks associated with ambient air toxics in US metropolitan areas. *Environmental Health Perspectives* 114(3):386–393.
19. Centers for Disease Control and Prevention (2004) Prevalence of diabetes among Hispanics–selected areas, 1998-2002. *Morbidity and Mortality Weekly Report* 53(40):941.
20. Centers for Disease Control and Prevention (2009) Differences in prevalence of obesity among black, white, and Hispanic adults-United States, 2006-2008. *Morbidity and Mortality Weekly Report* 58(27):740–744.
21. Sacco RL, Hauser W, Mohr J (1991) Hospitalized stroke in blacks and hispanics in northern Manhattan. *Stroke* 22(12):1491–1496.
22. White RO, Beech BM, Miller S (2009) Health care disparities and diabetes care: practical considerations for primary care providers. *Clinical Diabetes* 27(3):105–112.
23. US Bureau of Labor Statistics (2019) Employed persons by detailed industry, sex, race, and Hispanic or Latino ethnicity. *Labor Force Statistics from the Current Population Survey* https://www.bls.gov/cps/cpsaat18.htm.
24. Mollborn S, Fomby P, Dennis JA (2011) Who matters for children's early development? race/ethnicity and extended household structures in the United States. *Child Indicators Research* 4(3):389–411.
25. Bartel AP, et al. (2019) Racial and ethnic disparities in access to and use of paid family and medical leave: evidence from four nationally representative datasets. *Monthly Labor Review, U.S. Bureau of Labor Statistics*.
26. Small ML, McDermott M (2006) The presence of organizational resources in poor urban neighborhoods: An analysis of average and contextual effects. *Social Forces* 84(3):1697–1724.
27. Currarini S, Jackson MO, Pin P (2009) An economic model of friendship: Homophily, minorities, and segregation. *Econometrica* 77(4):1003–1045.
28. McPherson M, Smith-Lovin L, Cook JM (2001) Birds of a feather: Homophily in social networks. *Annual Review of Sociology* 27(1):415–444.
29. Belloni A, Chernozhukov V, Hansen C (2014) Inference on treatment effects after selection among high-dimensional controls. *The Review of Economic Studies* 81(2):608–650.
30. Belloni A, Chernozhukov V, Wei Y (2016) Post-selection inference for generalized linear models with many controls. *Journal of Business & Economic Statistics* 34(4):606–619.
31. Yang TC, Matthews SA (2015) Death by segregation: Does the dimension of racial segregation matter? *PloS One* 10(9).
32. Krivo LJ, Peterson RD, Kuhl DC (2009) Segregation, racial structure, and neighborhood violent crime. *American Journal of Sociology* 114(6):1765–1802.
33. Peterson RD, Krivo LJ (1993) Racial segregation and black urban homicide. *Social Forces* 71(4):1001–1026.
34. Shihadeh ES, Flynn N (1996) Segregation and crime: The effect of black social isolation on the rates of black urban violence. *Social Forces* 74(4):1325–1352.
35. Cooper RS (2001) Social inequality, ethnicity and cardiovascular disease. *International Journal of Epidemiology* 30(suppl_1):S48.
36. Morland K, Wing S, Roux AD, Poole C (2002) Neighborhood characteristics associated with the location of food stores and food service places. *American Journal of Preventive Medicine* 22(1):23–29.
37. Lincoln JR, Miller J (1979) Work and friendship ties in organizations: A comparative analysis of relation networks. *Administrative Science Quarterly* pp. 181–199.
38. Moody J (2001) Race, school integration, and friendship segregation in America. *American Journal of Sociology* 107(3):679–716.
39. Stearns E, Buchmann C, Bonneau K (2009) Interracial friendships in the transition to college: Do birds of a feather flock together once they leave the nest? *Sociology of Education* 82(2):173–195.
40. Wimmer A, Lewis K (2010) Beyond and below racial homophily: ERG models of a friendship network documented on Facebook. *American Journal of Sociology* 116(2):583–642.
41. Centers for Disease Control and Prevention (2020) Hospitalization rates and characteristics of patients hospitalized with laboratory-confirmed coronavirus disease 2019—COVID-NET, 14 states, March 1–30, 2020. *Morbidity and Mortality Weekly Report* 69.
42. Richardson S, et al. (2020) Presenting characteristics, comorbidities, and outcomes among 5700 patients hospitalized with COVID-19 in the New York City area. *Journal of the American Medical Association*.
43. Brayne S (2014) Surveillance and system avoidance: Criminal justice contact and institutional attachment. *American Sociological Review* 79(3):367–391.
44. Goffman A (2009) On the run: Wanted men in a Philadelphia ghetto. *American Sociological Review* 74(3):339–357.
45. Reardon SF, Firebaugh G (2002) Measures of multigroup segregation. *Sociological Methodology* 32(1):33–67.
46. CDC COVID-19 Response Team (2020) Preliminary estimates of the prevalence of selected underlying health conditions among patients with Coronavirus Disease 2019 — United States, February 12–March 28, 2020. *Morbidity and Mortality Weekly Report* 69:382–386.
47. Berkman LF (2009) Social epidemiology: Social determinants of health in the United States: Are we losing ground? *Annual Review of Public Health* 30:27–41.
48. Marmot M (2005) Social determinants of health inequalities. *The Lancet* 365(9464):1099–1104.
49. Acevedo-Garcia D, Lochner KA, Osypuk TL, Subramanian SV (2003) Future directions in residential segregation and health research: A multilevel approach. *American Journal of Public Health* 93(2):215–221.
50. Ellen IG, Mijanovich T, Dillman KN (2001) Neighborhood effects on health: Exploring the links and assessing the evidence. *Journal of Urban Affairs* 23(3-4):391–408.
51. Frank KA (2000) Impact of a confounding variable on a regression coefficient. *Sociological Methods & Research* 29(2):147–194.
52. Oster E (2019) Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics* 37(2):187–204.
53. Bolin B, Kurtz LC (2018) Race, class, ethnicity, and disaster vulnerability in *Handbook of Disaster Research*. (Springer), pp. 181–203.
54. Lariscy JT, Hummer RA, Hayward MD (2015) Hispanic older adult mortality in the united states: New estimates and an assessment of factors shaping the hispanic paradox. *Demography* 52(1):1–14.
55. Kermack WO, McKendrick AG (1927) A contribution to the mathematical theory of epidemics. *Proceedings of the Royal Society of London* 115(772):700–721.
56. Tibshirani R (1996) Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)* 58(1):267–288.