Submitted to Econometrica

A Novel Solution to Biased Data in Covid-19 Incidence Studies

H. D. Vinod and Katherine Theiss

June 28, 2020

Submitted to Econometrica

A NOVEL SOLUTION TO BIASED DATA IN COVID-19 INCIDENCE STUDIES

H. D. Vinod and and Katherine Theiss

Complete novelty and uncertainty of the Covid-19 pandemic have created many challenging scientific problems, including biased data arising from a lack of randomized testing over the general population. We describe the bias problem and its solution from Econometrics literature, which seems to have been neglected by epidemiology experts. We study a large Covid-19 US data set, providing nationwide forecasts of deaths to illustrate the model's power. Our two-equation model overcomes the bias by using the inverse Mills ratio and improves forecasts of new deaths in all nine out of nine weekly out-of-sample comparisons. It can be applied to a variety of problems associated with the pandemic. A focused study of trends in deaths predicted by lagged cumulative infections reveals that forty-two states have negative trends and that seven of nine states with undesirable positive trends have Republican governors.

Keywords: Inverse Mills Ratio, Selection Models, Poisson probit model, outcome equation.

1. AN EXTENDED GENERALIZED LINEAR MODEL

The Covid-19 pandemic has had a profound effect on many aspects of life around the world. We focus on a statistical component of prediction and mitigation of any undesirable outcome Y_{it} at time t. For an illustration of our newer methodology, our Y_{it} is the count of deaths from Covid-19 in a state-level jurisdiction (50 states plus the District of Columbia or DC). It is convenient to refer to the mouthful 'state-level jurisdictions,' simply as 'states,' after inserting the quotation marks if needed.

Our panel (longitudinal) data are described later in detail in Section 3. Briefly, our data consists of various variables for 51 'states' with daily details over 90 days in the year 2020. We note that many variables do not change at

Professor of Economics, Fordham University, 441 E Fordham Rd, Bronx, NY 10458.
Ph.D. Candidate at Fordham University, 441 E Fordham Rd, Bronx, NY 10458.

all from day to day. The outcomes studied in our empirical illustration are Y_{it} as counts of deaths in 'state' numbered $i = 1, \ldots, 51$ on day $t = 1, \ldots, 90$.

The mathematical model for infectious diseases from the 1920s, Kermack and McKendrick (1991), divides the surviving population (after subtracting the deaths) into three groups for susceptible, infected and recovered (SIR). Many epidemiological models are used to study the impact of Covid-19 which focus on a cumulative sum of deaths and rates of change in the cumulative totals. This paper focuses on a unique data limitation arising from early unavailability of reliable tests for Covid-19 and using econometric tools to overcome the limitation. We are unaware of any epidemiological model for Covid-19, which uses the 'inverse Mills ratio,' suggested here.

The most common technique employed to model count data is Poisson regression. Thus, we use a Generalized Linear Model (GLM) with the Poisson distribution providing a so-called 'link function' to predict Covid-19 deaths based on the cumulative infections from the previous week, $X_{i,t-1}$. More specifically, we regress the cumulative death count Y_{it} on the log of lagged cumulative infection count $X_{i,t-1}$, where time is indexed by weeks.

A simplified version of our outcome regression is:

$$Y_{it} = \beta_0 + \beta_1 log(X_{i,t-1}) + \epsilon_{it}, \tag{1}$$

where ϵ_{it} denotes errors.

Our model (1) is fairly general in the sense that it can represent other undesirable outcomes, such as hospitalizations, costs, or losses, with appropriately chosen control variables and link functions (if any). Before we can estimate the model, we must recognize a data problem described in the next section, eventually leading to our solution using the 'inverse Mills ratio.'

2. DATA PROBLEMS IN STUDYING COVID-19

The Covid-19 virus first appeared around December 2019 in China. Since it was novel, initially no one had a safe and effective diagnostic test or treatment, and certainly no vaccine. It was known to be highly contagious, but the mode of transmission and its deadliness was uncertain.

In early months of the outbreak, it was most important to know if a sick person being admitted to a hospital was infected by Covid-19. Since the virus is brand new, reliable tests and lists of all symptoms were slow, expensive, and hard to find.

How is the limited supply of tests rationed? We suspect that the priority was given mostly to the following persons in the following order.

(a) patients at the hospital door showing severe symptoms (whose list is evolving and expanding),
(b) the front-line doctors and nurses having direct contact with infected patients,
(c) support staff having direct contact with infected patients,
(d) all other doctors, nurses and support staff working in the same facility (hospital or nursing home) as infected patients,
(e) persons whose primary physicians suspected that they have the virus and prescribed a test,
(f) persons who traveled from infected areas,
(g) employees of essential businesses (e.g., food delivery), and
(h) persons who suspect that they were exposed to an infected person.

The limited supply of tests had to be allocated to these individuals before they could be used on a random sample from the general population. If we want to draw statistical inference about the entire population, it is not appropriate to use a biased sample of data only coming from a small subgroup of the population with unique characteristics. The unknown bias is related to the dependence of health outcomes on (a) to (h).

2.1. *Probit model for the probability of being tested*

Direct data on items (a) to (h) above is not available to us and generally not collected. We find that certain state-level data are available, which can proxy these items. Our approach is inspired by Heckman (1979) who also suggests explicitly modeling a "choice equation" for estimating the probability of being chosen to be in the data. Our choice of the variable is being chosen (or allowed) by authorities to be tested for Covid-19. We can write the choice equation as a linear regression equation in a compact notation using long vectors and large matrices by single symbols as:

$$C^* = Z\gamma + \delta \quad \text{(choice equation)}, \tag{2}$$

where C^* is the unobserved propensity to be selected in the special group of people who are tested for Covid-19, Z denotes a matrix of data on observable explanatory variables, γ denotes regression coefficients, and $\delta \sim N(0,1)$ is the error distributed as a standard normal random variable with mean zero and variance unity.

Our empirical model uses the following 'state' level data as columns in our Z matrix. We report their abbreviations in parentheses:

(i) hospital employee share (HES),
(ii) the proportion of people that commute on public transit (CPT),
(iii) hypertension rate (HR), which has been one of the leading comorbidities,
(iv) the proportion of uninsured (UI) population, and
(v) median household income (HI).

The asterisk in the notation C^* in (2) suggests that propensity cannot be directly measured. The notation C refers to the measurable binary variable, which takes only two values. $C = 1$ only when the underlying probability of being chosen is positive, and $C = 0$ measures the probability of not being

chosen for testing.

The observable counterpart of (2) is

$$C_{it} = \gamma_0 + \gamma_1 HES_{it} + \gamma_2 CPT_{it} + \gamma_3 HR_{it} + \gamma_4 HI_{it} + \gamma_5 UI_{it} + \delta_{it} \tag{3}$$

We expect that a randomly chosen person from 'states' with higher values of HES, CPT, HR, and HI is more likely to be tested. By contrast, a person from a 'state' with a large number of people lacking health insurance (high UI) is relatively *less* likely to be chosen for Covid-19 testing. That is, we expect $\gamma_j > 0, j = 1, 2, 3, 4$, and $\gamma_5 < 0$.

2.1.1. *Binary data attribution at the individual patient level*

We do not have data on individual patients. Yet, the assessment of the probability of being tested for Covid-19 requires an approximation to such data based on available information. We use stratified random sampling to generate simulated data on 500,000 individuals. Addressing these individuals in the formulas requires an additional subscript k representing an individual who is either tested for Covid-19 or not. We do not display the k subscript to avoid cluttered notation in (3).

We assigned each individual to a particular 'state,' based on the weighted probability of residence. This weighted probability was calculated by taking each state's total population and dividing by the entire United States population. Then, each individual was assigned a testing index of 1 or 0 based on the weighted probability that they were tested for Covid-19 in their assigned state. We calculate the weighted probability by taking the cumulative number of tests administered in each state and dividing by the total state population. We assigned socioeconomic and demographic variables to the individual based on his state of residence. We repeat this process for each week between 2020/04/20 and 2020/06/15.

The regression model (3) has several unique features discussed in Econometrics textbooks, (Vinod, 2008, ch. 7). The dependent variable is said to be binary, since $C_{it} = 1$ or $C_{it} = 0$ are the only two values observed. Hence the conditional expectation $E(C_{it}|Z) = Z\gamma$, allows only two values for the error term: $\delta = 1 - Z\gamma$, or $\delta = Z\gamma - 1$, to ensure that they add up to zero, and satisfy $E(\delta) = 0$, since we have assumed that $\delta \sim N(0, 1)$.

We want to model the probability of being tested for Covid-19 using data on the right-hand side, RHS=(HES, CPT, HR, UI, HI). Clearly, we need to use some transformation of variables to ensure that that all fitted values for the probability of 'being chosen to be tested' lie in the [0,1] range. We know that the cumulative density of the standard normal density, $\Phi(z)$, is also in the same range. The probit model exploits the property $\Phi(z) in [0, 1]$ to represent the probability of being tested for Covid-19 on the left-hand side. However, we cannot expect the explanatory RHS variables to also lie in the [0,1] range. The beauty of the probit model is that it avoids imposing such unrealistic demands of the values of RHS variables.

The probit regression model rewrites (3) by replacing the binary dependent variable on the left-hand side by the probability of 'being chosen to be selected for virus testing' conditional on a given value of right-hand side variables Z_{it}, as the regression:

$$P(C_{it} = 1 \mid Z_{it}) = \Phi(\mathrm{RHS}_{it}) + \delta_{it}, \tag{4}$$

where the range of values of the dependent variable is restricted to the [0,1] range, even if the RHS values are anywhere on the real line in $(-\infty, \infty)$. The probit specification guarantees that the fitted values will also lie in the [0,1] range. We use the R software function "glm" for this purpose. The biometric terminology calls $\Phi(\mathrm{RHS}_i)$ as a link function $h(Z\gamma)$.

By analogy with the choice equation (2), the outcome equation is

$$Y^* = X\beta + \varepsilon \tag{5}$$

in matrix notation with the asterisk in Y^* to remind us that the outcome data is not available for the general population, but only for an unrepresentative (biased) sample of the cases when we perform the diagnostic test.

To repeat, the available sample of tested people is obviously not randomly chosen from the general population, and yet we want to model the Covid-19 impact on the general population and make statistical inferences about the outcome. The bias arises because the selection criterion of the choice equation (2) is correlated with the dependent variable of the outcome equation (5). In the present case, the outcome represents death from Covid-19.

In the outcome regression, we are not estimating Y, the true projected death count for the whole population, because it is not observed due to data limitations. We can regard Y^* as only a "preliminary" indicator of health outcome for the general population facing the novel virus.

The prediction of deaths Y^* is only based on known infections reported from those who were tested ($C^* > 0$) for the virus. Formally, we use a mathematical device called the indicator function $I_{nd}(w)$, where w is some condition, which is true or false. The indicator function has the value unity only if the condition w is true, and zero otherwise.

$$Y = Y^* I_{nd}(C^* > 0), \quad Y = 0 \ \text{ if } C^* \leq 0 \ \text{ and } \ Y = Y^* \ \text{ if } C^* > 0. \tag{6}$$

Clearly, the preliminary health outcome Y^* will equal the true health outcome Y only for the subset of the population who are tested for Covid-19. The fact that the outcome Y is measurable only for those who are tested is expressed formally in eq. (6) as $Y = Y^*$ if $C^* > 0$.

The selection bias arises if the error in the choice equation δ is correlated

with the error in the outcome equation. That is, if the choice to be tested for the virus is correlated with the (unobservable part of the) outcome of interest.

According to the classic solution to the problem, Heckman (1979), it is assumed that δ and ε are jointly normally distributed:

$$\begin{pmatrix} \varepsilon \\ \delta \end{pmatrix} = N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \sigma_\varepsilon & \sigma_{\delta,\varepsilon} \\ \sigma_{\delta,\varepsilon} & 1 \end{pmatrix}\right). \tag{7}$$

Recall that the definition $C^* = Z\gamma + \delta$ means that $\delta = C^* - Z\gamma$. The variance of δ is unknown and assumed to be unity. Under this assumption, the regression function for the observed dependent variable Y can be written as:

$$E(Y \mid X) = E(Y^* \mid X, C^* > 0) \tag{8}$$

$$= E(X\beta \mid X, C^* > 0) + E(\epsilon \mid X, C^* > 0) \tag{9}$$

$$= X\beta + E(\epsilon \mid X, Z\gamma > -\delta) \tag{10}$$

$$= X\beta + \sigma_{\delta\epsilon}\left[\frac{\phi(Z\gamma)}{\Phi(Z\gamma)}\right], \tag{11}$$

where ϕ is the Normal density and Φ is the cumulative density of the Normal, as before, and the ratio $[\phi(Z\gamma)/\Phi(Z\gamma)]$ is known as the "inverse Mills ratio," (IMR), based on the well-known properties of the truncated Normal density. The coefficient of IMR is seen in eq. (11) to be $\sigma_{\delta\epsilon}$, the covariance between δ and ϵ. Typical Econometrics texts, (Vinod, 2008, Sec. 7.3), estimate parameters β, γ, and the covariance by a two-step procedure.

The use of IMR to reduce selection bias is popular in Econometrics, and Professor Heckman had received the Nobel prize in Economics for inventing it. Its use in bias-reduction in Covid-19 studies is promoted here. The probit step estimates γ by attempting to estimate the probability of being chosen to be tested for Covid-19, i.e., having $C^* > 0$. It is modeled as a probability

of an indicator function I_{nd} being positive:

$$P[I_{nd}(C^* > 0 \,|\, Z) = 1] = \Phi(Z\gamma). \tag{12}$$

Once the estimate of γ is available, it can be substituted in the correction term $[\phi(Z\gamma)/\Phi(Z\gamma)]$ known as the inverse Mills ratio. Using only observations with $I_{nd}(C^* > 0 \,|\, Z) = 1$, we can make the estimated inverse Mills ratio as a second regressor in the Poisson outcome equation (1) and estimate $\sigma_{\delta\epsilon}$ as the regression coefficient in light of eq. (11).

Now our outcome equation (1) becomes:

$$Y_{it} = \beta_0 + \beta_1 log(X_{i,t-1}) + \beta_2(IMR_{it}) + \epsilon_{it}, \tag{13}$$

The corresponding fitted values are given by using the 'hat' symbols to denote estimated values as:

$$\hat{Y}_{it}^{imr} = \hat{\beta}_0 + \hat{\beta}_1 log(X_{i,t-1}) + \hat{\beta}_2(IMR_{it}), \tag{14}$$

where we have a superscript 'imr' when the regressor IMR is present. We use an R package Toomet and Henningsen (2008) for estimating IMR.

We study the usefulness of IMR by considering an alternate model which does not have IMR, having the fitted values denoted without the superscript as:

$$\hat{Y}_{it} = \hat{\beta}_0 + \hat{\beta}_1 log(X_{i,t-1}). \tag{15}$$

We remind the reader that these equations have an additional subscript k to represent the individual person who is tested for the virus. We have been avoiding the additional subscript as in (3) to avoid clutter. It is understood that the fitted values $\hat{Y}_{it}^{imr}, \hat{Y}_{it}$ represent averages over the k subscript for

individuals in the data.

3. DETAILS OF DATA SOURCES

Table I describes the sources of data. Further details about the items in column entitled 'Source' are as follows. Data on Covid-19 testing indicator, ($C_{it} = 0, 1$, at time t for state i) infection numbers and deaths were obtained from the Covid Tracking Project (CTP) (`https://covidtracking.com`). The two character codes for various 'states' are from the US postal service (USPS) (`https://www.usps.com/`).

The population in 2019 (POP) is from (`https://www.census.gov/data/tables/time-series/ demo/popest/2010s-state-total.html`). All relevant socioeconomic and demographic variables were obtained from the United Census Bureau (UCB) (`https://www.census.gov`), the Bureau of Labor Statistics (BLS),(`http://www.bls.gov`), American Health Rankings (AHR), (`https://www.americashealthrankings.org`) the Bureau of Transportation Statistics (BTS), (`https://www.bts.gov`) and World Population Review (WPR) (`https://worldpopulationreview.com/`).

TABLE I
DESCRIPTION OF DATA AND THEIR CODES

	Code	Description	Source
1	C_{it}	testing indicator	CTP
2	STA	State Abbreviation	USPS
3	POP	Population in 2019	UCB
4	HES	Hospital Employee Share	BLS
5	UI	Uninsured Population	AHR
6	HR	Hypertension Rate	AHR
7	CPT	Commute Public Transit	BTS
8	HI	Household Income	WPR

3.1. *Grouping states by political affiliations*

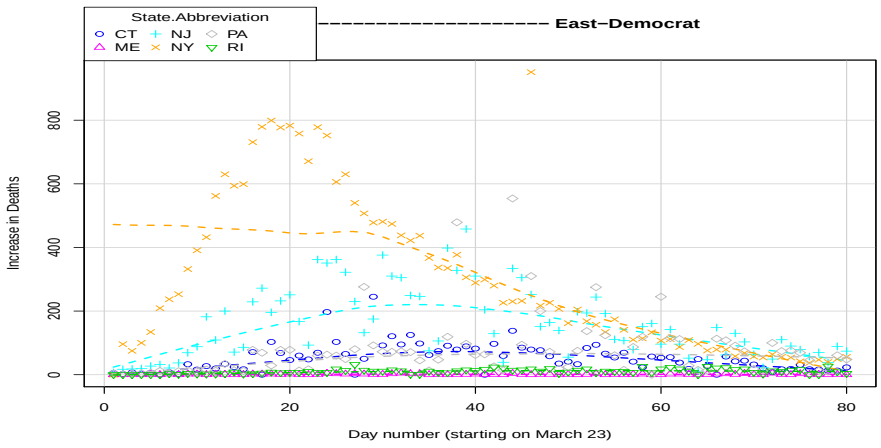
In this section, we report graphs of our data by states in two sets. First, as scatterplots to depict heterogeneity across states and second as heterogeneity across time intervals. Instead of an alphabetic grouping of states, it seems to be more interesting to divide the states into the following groups based on region and politics.

TABLE II

GROUPING OF STATES BY REGION AND PARTY AFFILIATION OF ITS GOVERNOR

j	Short Name	Region	Governor Affiliation	Postal service state codes
1	EDem	East	Democrat	CT, ME, NJ, NY, PA, RI
2	ERep	East	Republican	AL, FL, MA, MD, NH, SC, VT
3	SDem	South	Democrat	DC, DE, KY, LA, NC, VA
4	SRep	South	Republican	AR, GA, MS, OK, TN, TX, WV
5	WDem	West	Democrat	CA, CO, HI, MT, NM, NV, OR, WA
6	WRep	West	Republican	AK, AZ, ID, UT, WY
7	MidWDem	Midwest	Democrat	IL, KS, MI, MN, WI
8	MidWRep	Midwest	Republican	IA, IN, MO, ND, NE, OH, SD

Figure 1: Scatterplot of data for Region 1



Figures 1 to 8 plot our data for each day starting from March 25, 2020, for a total of 80 days. The regions are $j = 1, \ldots 8$ as listed in the first

Figure 2: Scatterplot of data for Region 2

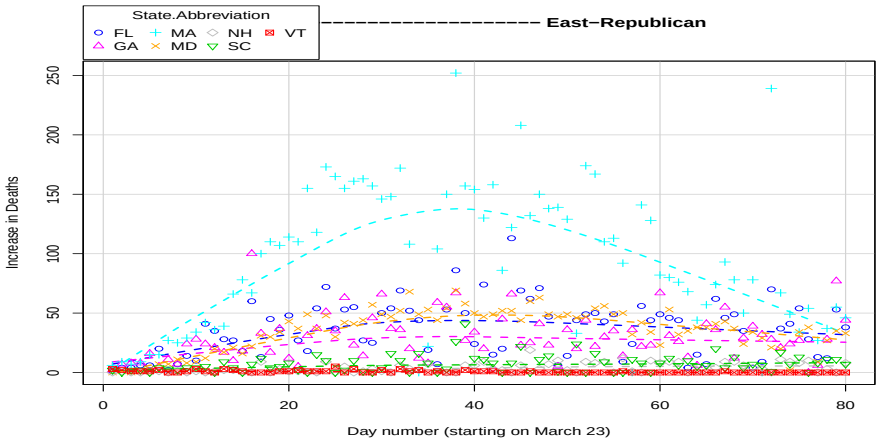
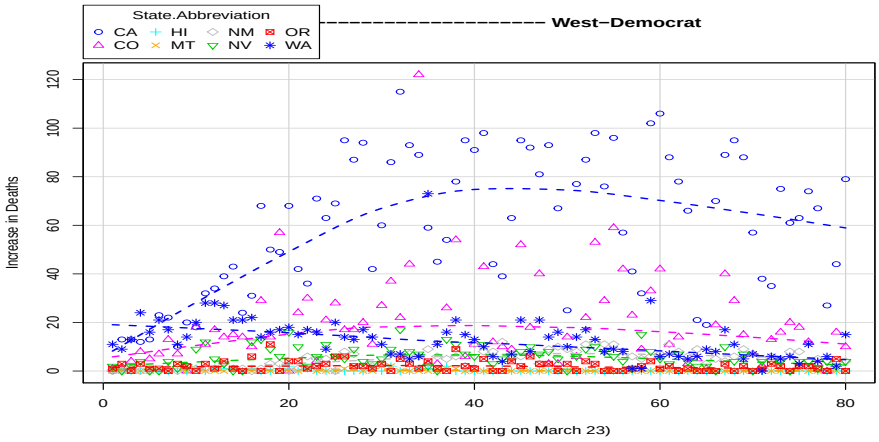


Figure 3: Scatterplot of data for Region 3



column of Table II. We have produced eight plots depicting the day-to-day heterogeneity for each region. For brevity we report only one in Figure 9.

As testing for the Covid-19 infections becomes more readily available, we have a larger count of infected people showing positive test results. It is of interest to know whether the ratio of (newly infected) to (newly tested) persons is declining over the time period studied here. Figures 10 to 17 plot our data for each day starting from March 25, 2020, for a total of 80 days.

Figure 4: Scatterplot of data for Region 4

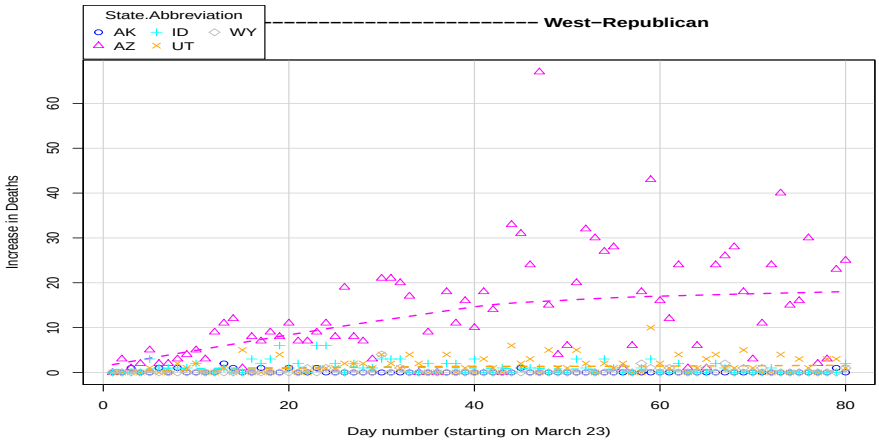
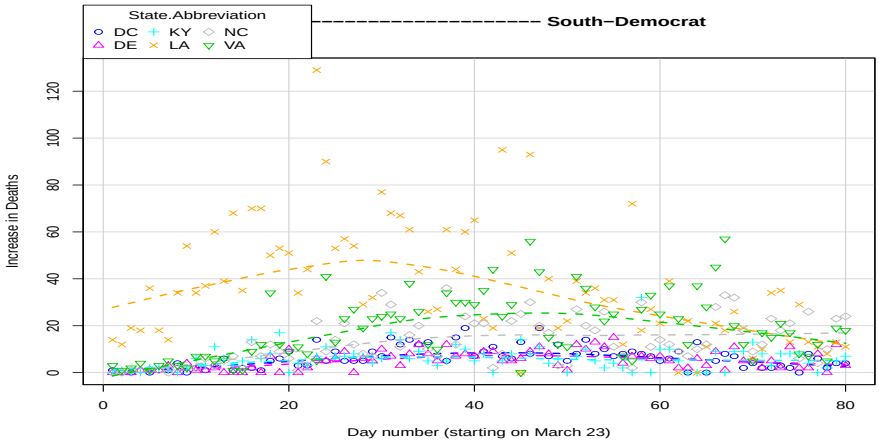


Figure 5: Scatterplot of data for Region 5



The regions are $j = 1, \ldots 8$ as listed in the first column of Table II. We have produced eight plots depicting the day-to-day heterogeneity for each region. Again, for brevity, we report only one in Figure 18.

Figure 6: Scatterplot of data for Region 6

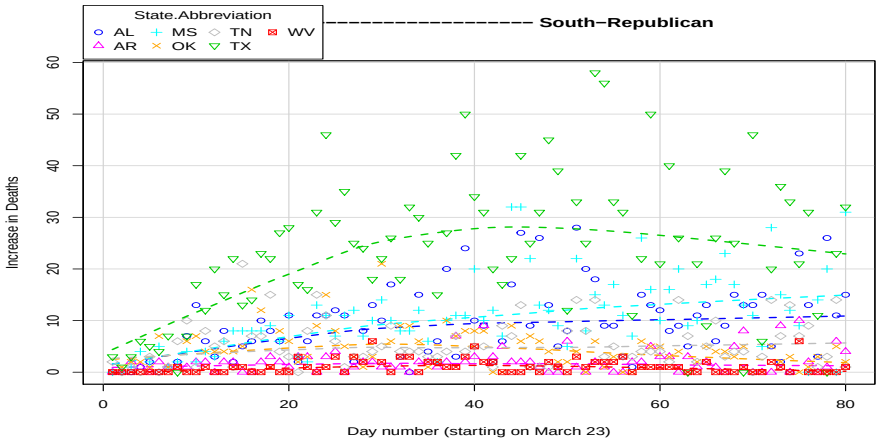


Figure 7: Scatterplot of data for Region 7

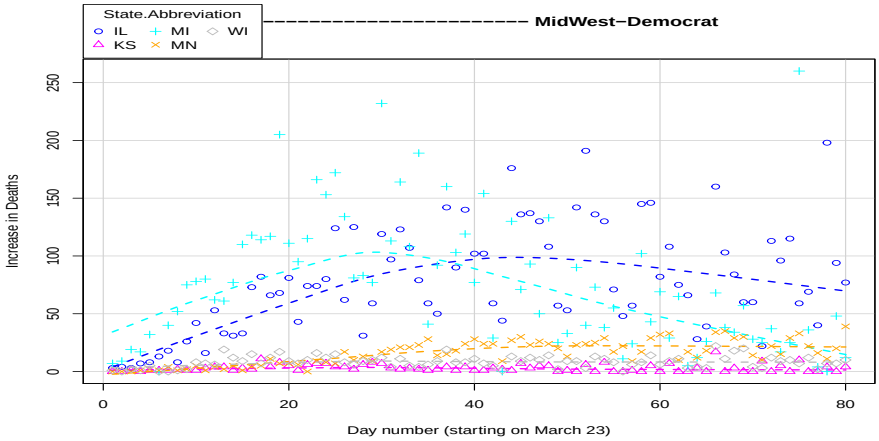


Figure 8: Scatterplot of data for Region 8

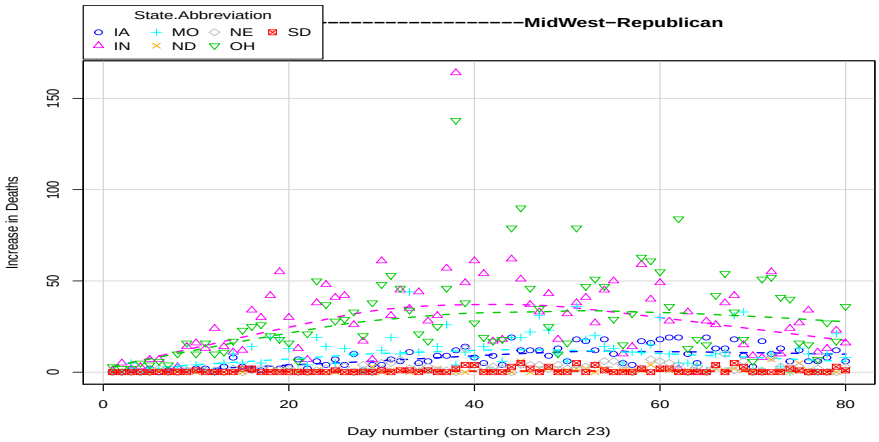


Figure 9: Typical day-by-day heterogeneity of data for Region 1

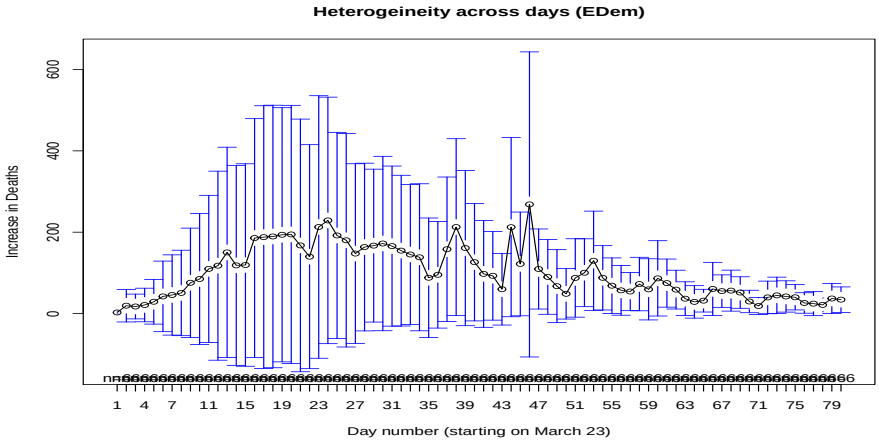


Figure 10: Scatterplot of the ratio of newly infected to newly tested for Covid-19 for Region 1

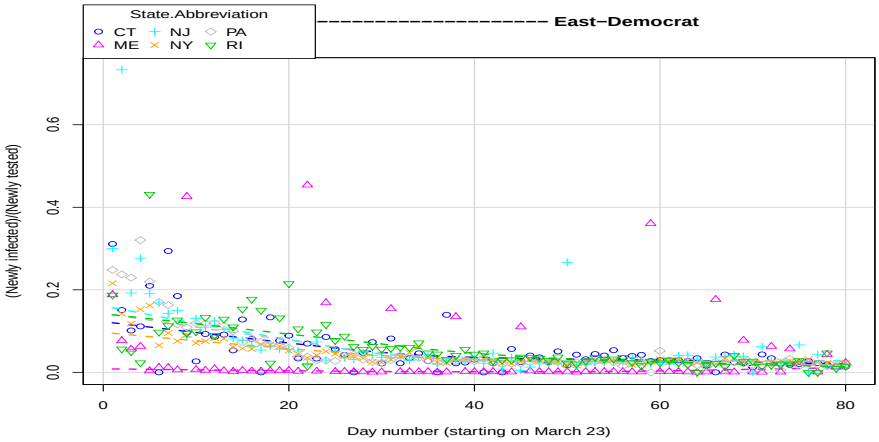


Figure 11: Scatterplot of the ratio of newly infected to newly tested for Covid-19 for Region 2

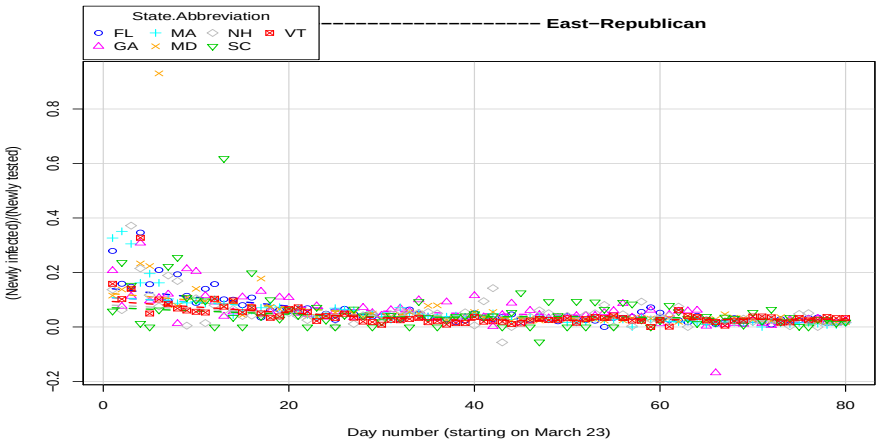


Figure 12: Scatterplot of the ratio of newly infected to newly tested for Covid-19 for Region 3

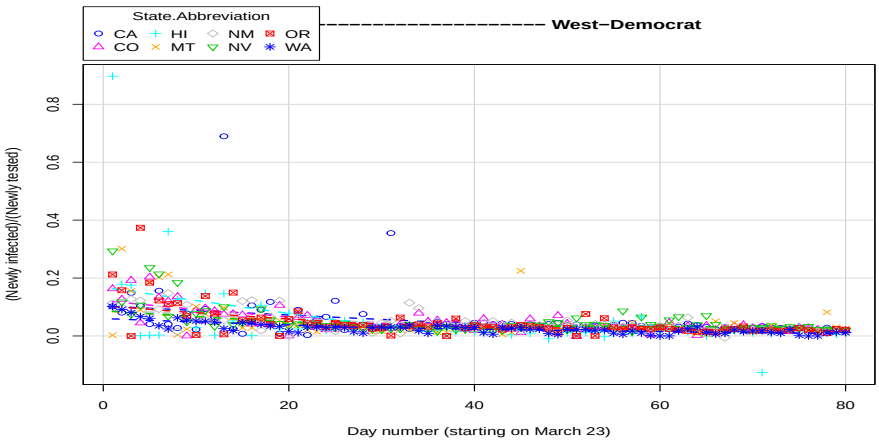


Figure 13: Scatterplot of the ratio of newly infected to newly tested for Covid-19 for Region 4

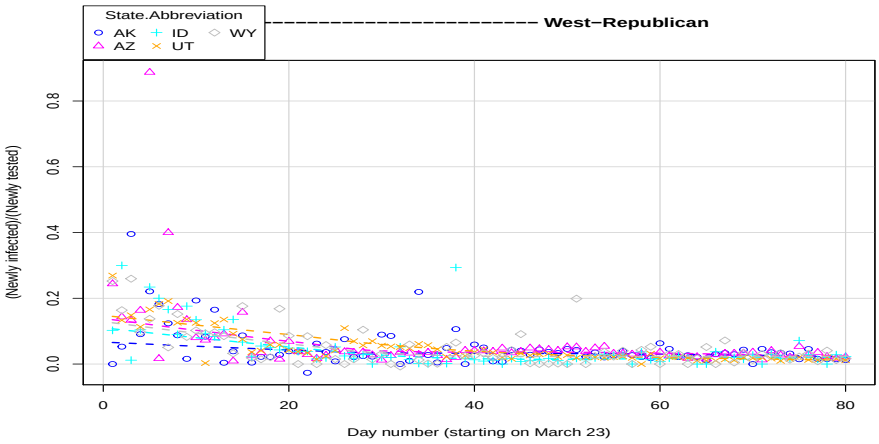


Figure 14: Scatterplot of the ratio of newly infected to newly tested for Covid-19 for Region 5

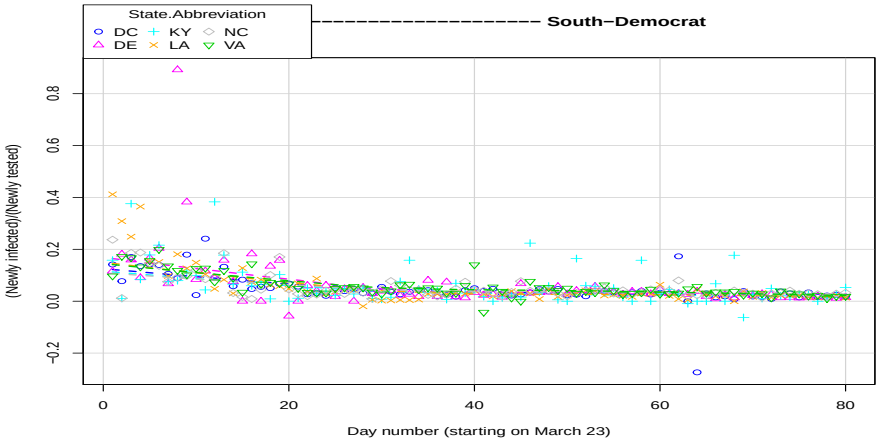


Figure 15: Scatterplot of the ratio of newly infected to newly tested for Covid-19 for Region 6

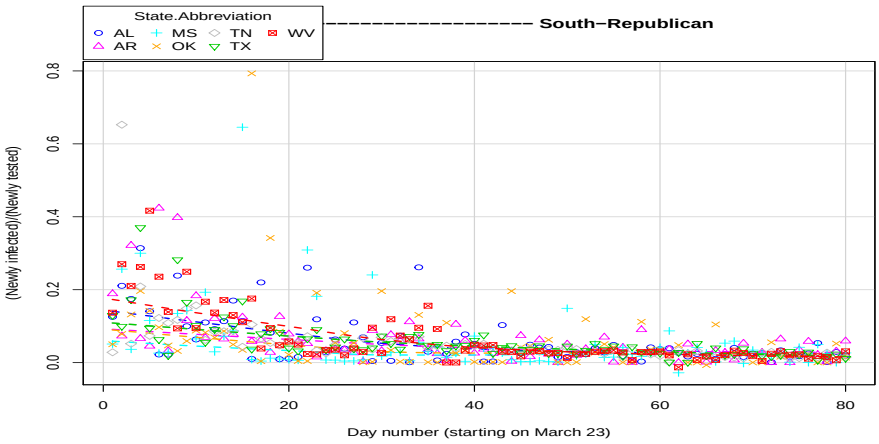


Figure 16: Scatterplot of the ratio of newly infected to newly tested for Covid-19 for Region 7

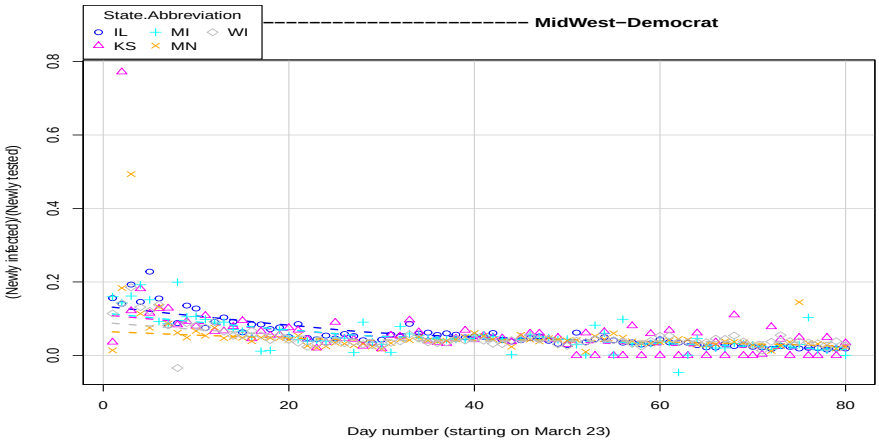


Figure 17: Scatterplot of the ratio of newly infected to newly tested forCovid-19 for Region 8

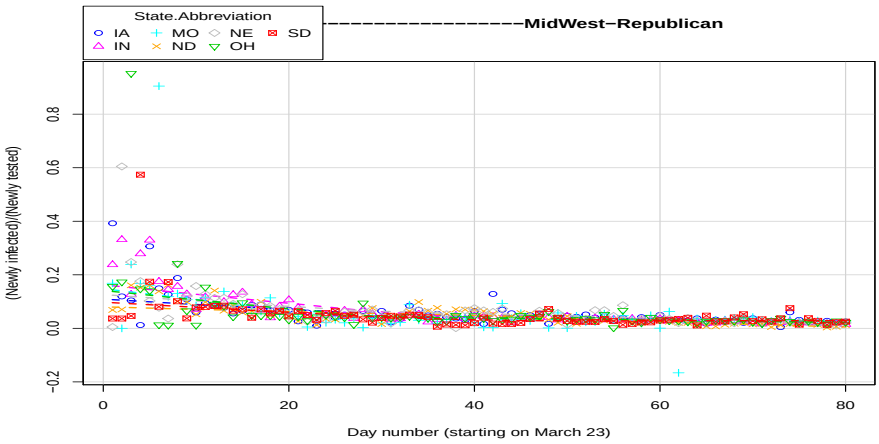
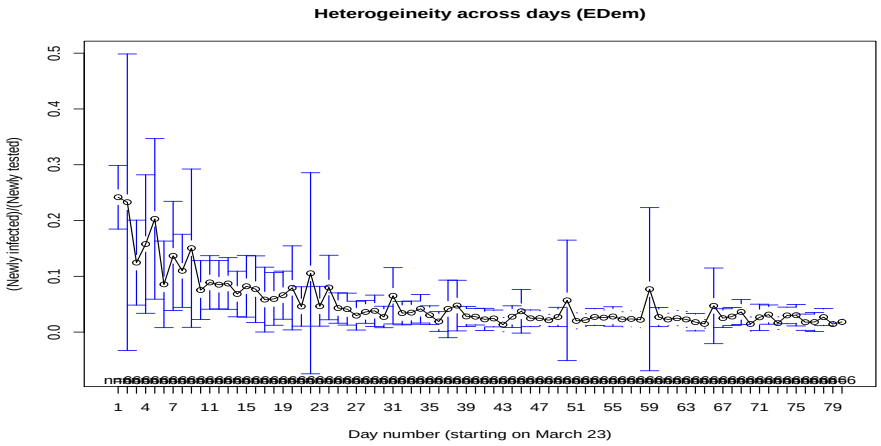


Figure 18: Typical day-by-day heterogeneity of the ratio of newly infected to newly tested for Covid-19 for Region 1



4. A FOCUSED STUDY OF TRENDS IN ρ OVER 11 WEEKS

We are interested in seeing whether the mitigation strategy has worked by producing a declining trend in new deaths (ND) in a sense defined here. We hope that our evaluation provides a distinct insight supplementing the notion of flattening the curve of deaths used by epidemiologists.

Consider a simplified version of outcome equation (1) where we regress new deaths ND_{it} in-state i and week (t) on the one-week lagged value of the cumulative number of infections (CI). We do not use logs, and our weekly data set begins with the week starting on March 22 as the first week $(t = 1)$ when the mitigation was started. Our data set for this section ends with the week starting on May 31 as our 11-th week $(t = 11)$. Also, unlike (1), we do not include any additional explanatory variables. We also do not directly adjust for selection bias using the IMR variable as a regressor, as in (13). Our simplified outcome regression for a study of trends in new deaths is:

$$ND_{it} = \beta_0 + \rho\,(CI_{i,t-1}) + \epsilon_{it}, \tag{16}$$

where ϵ_{it} denotes errors, the key parameter of interest is the coefficient ρ of a lagged variable.

We can use (16) for prediction of Covid-19 deaths from the estimate of β_0, and the estimate $\hat{\rho}$ times the cumulative infections of the previous week, $(CI_{i,t-1})$.

Assessing Changing Virus Damage from Estimates of ρ

The interpretation of the estimate $\hat{\rho}$ in eq. (1) is as follows.

(I) If $\hat{\rho}$ increases from week to week, the new damage from Covid-19 is increasing over time, making it a cause of concern.

(II) If $\hat{\rho}$ remains the same from week to week, the new damage from Covid-19 is steady over time, suggesting greater progress might be

needed for further reductions.

(III) If $\hat{\rho}$ decreases from week to week, the new damage from Covid-19 is decreasing over time, suggesting good news.

As time passes, new data are made available, and each new data on (Y_{t+1}, X_{t+1}) will create a new estimate of ρ, which may be denoted as $\hat{\rho}_{t+1}$. This will create over time a set of estimates of $\hat{\rho}_{t+j}, j = 0, 1, \ldots, h$ providing a somewhat new way of studying the damage from Covid-19. It is claimed to be new, since we are not aware of any Covid-19 study focusing on the predictive parameter ρ.

We can find eleven estimates of changing ρ values defined in equation (16) as the coefficient of the lagged dependent variable. These values indicate the progression of the estimates over time.

Let us define $\tau = \{1, 2, \ldots 11\}$ and further define the following quadratic regression in generic notation.

$$\hat{\rho} = a + b\tau + c\tau^2 + u \tag{17}$$

We have estimates for 51 'states' for this equation in two versions, one linear by assuming that $c = 0$, and another as in (17). Instead of reporting 51 trends, we used the grouping of states defined in Section 3.1. We want to know if there is a positive or negative trend in these $\hat{\rho}$ values. In the case of a linear model, (with c absent), the sign of the trend is simply the sign of the estimated slope $\hat{b}$.

In the quadratic case, the estimated derivatives for each state are

$$\frac{\partial \hat{\rho}}{\partial \tau} = \hat{b} + \hat{c}\tau. \tag{18}$$

Instead of working with eleven estimates of these partials at each value of τ, it is convenient to summarize them by evaluating the right-hand side of (18) at the sample mean of the regressor $\bar{\tau} = 6$. If $\partial\hat{\rho}/\partial\tau$ evaluated at $\bar{\tau} = 6$

TABLE III

REGIONAL SUMMARY OF ρ TRENDS

	i	n+	n-	Signif	Insignif	qu+	qu-	cor
EDem	1	1	5	1	5	0	6	-0.2330
ERep	2	2	5	1	6	2	5	0.3579
SDem	3	1	5	0	6	0	6	0.1563
SRep	4	1	6	2	5	1	6	0.5955
WDem	5	3	5	1	7	2	6	-0.1643
WRep	6	3	2	1	4	2	3	-0.2157
MidWDem	7	0	5	2	3	0	5	-0.3229
MidWRep	8	3	4	1	6	2	5	-0.6152
sum	–	14	37	9	42	9	42	-0.4415
average	–	2	5	1	5	1	5	-0.0552

is negative, we have a desirable decline in the outcome from Covid-19.

Table III reports a summary of our estimated trend directions for each of the eight regions named in Table II. The column titles mean the following. "n+" reports the count of the number of positive values in that region. "n-" reports the count of the number of negative values in that region. "Signif" reports the count of the number of (small, < 0.05) p-values suggesting statistical significance of the trend in that region. "Insignif" reports the count of the number of (large) p-values suggesting statistically insignificant trends in that region. "qu+" reports the count of the number of positive values of $(\partial\hat{\rho}/\partial\tau > 0)$, in that region using the quadratic formulation (17). "qu-" reports the count of the number of negative values of $(\partial\hat{\rho}/\partial\tau < 0)$ in that region using the quadratic (17). "cor" reports the simple correlation between the bias measure associated with each state based on inverse Mills ratio (IMR) and slope of the linear trend within the particular group of states.

The bottom two lines report the column sum and column mean rounded to zero decimals. The row marked "sum" suggests that of the 51 states 37 had negative linear trends, and 14 had positive trends. The vast majority of the trend coefficients are statistically insignificant. This means that the

observed negative overall trends may well turn positive.

The row "sum" for column "qu+" has a nine. It means the following. The quadratic fit is more realistic than linear, and partials are mostly negative in 42 states and positive in nine states. Of the nine positively trending states, two each are from MidWRep, WRep, and ERep regions and one from SRep, showing that seven of nine states with positive trends have Republican governors.

The correlation coefficient between trend slopes and IMR are negative in four of eight groups of states, suggesting that as the bias index IMR increases, the trend decreases. Three groups of states (SRep, ERep, and SDem) have positive correlations, implying that higher bias is associated with higher slopes of linear trends within those states.

5. OUT-OF-SAMPLE NATION-WIDE FORECAST ERRORS: MODELS WITH AND WITHOUT IMR

We run the Poisson model with and without the inverse Mills ratio over nine weekly values for each state starting with $t = 1$ for the week starting on April 20, 2020, and ending with $t = 9$ for the week starting on June 15, 2020. We find that the inverse Mills ratio is highly significant in regressions (13) during each time period.

Now we want to report estimation results for two versions of fitted values for each state i and week number t defined in equations (14) and (15) as $\hat{Y}_{it}^{imr}, \hat{Y}_{it}$, respectively, for two models, with and without IMR, respectively. One calls them out-of-sample predictions because they predict deaths during the next week defined as the week $(t+1)$ based on information through week t only. The estimation process does not access any information available only at time $(t + 1)$.

As researchers, we have access to the actual death counts at time $(t + 1)$ for each state $Y_{i,t+1}$. Since the sizes of various states in the US differ a great deal, it is not appropriate to compare death count numbers them-

selves with corresponding predictions $\hat{Y}_{i,t+1}$. A better metric for comparison is the percent error. Instead of forecast errors for individual states, this section reports nationwide results. Our speculative implications for state-wise infections are discussed later in Section 6.

Let $TAD = \sum_{i=1}^{51} Y_{i,t+1}$ denote nation-wide total actual deaths by letting the outcome Y represent new deaths (ND) by adding over all states. Let TPD denote country-wide total predicted deaths, and let the outcome Y be new deaths (ND). Now we have:

$$TPD_{t+1} = \sum_{i=1}^{51} \hat{Y}_{it}, \qquad TPD^{imr}_{t+1} = \sum_{i=1}^{51} \hat{Y}^{imr}_{it}, \tag{19}$$

In general, the percent forecast error between any predicted values (PV) and corresponding actual values (AV) is $\%\Delta(PV, AV) = 100(PV - AV)/AV$. In our case, we are subtracting the actual nation-wide deaths from their predicted values, as opposed to $(AV - PV)$, so that underestimated forecasts of deaths become intuitively meaningful negative values. Results from all weeks are reported in Table IV, where we use the definition of percent underestimation as:

$$\%\Delta_{t+1}(TPD, TAD) = 100\,\frac{(TPD_{t+1} - TAD_{t+1})}{TAD_{t+1}}, \tag{20}$$

for a model without IMR, and analogously defined $\%\Delta_{t+1}(TPD^{imr}, TAD)$ for a model with IMR.

Figure 19 plots the country-wide results from table IV. We observe that the Poisson model with the IMR more accurately predicts the death count than does the Poisson model without the IMR along every row of the Table. The percent error is numerically larger for all weeks for models without the IMR bias correction. The improvement is measured out-of-sample by comparing the actual death counts predicted for the week t based only on

TABLE IV

PERCENT UNDERESTIMATION IN OUT-OF-SAMPLE FORECASTS OF DEATHS USING EQ. (20) WITH AND WITHOUT IMR FOR WEEKS STARTING ON INDICATED DATES.

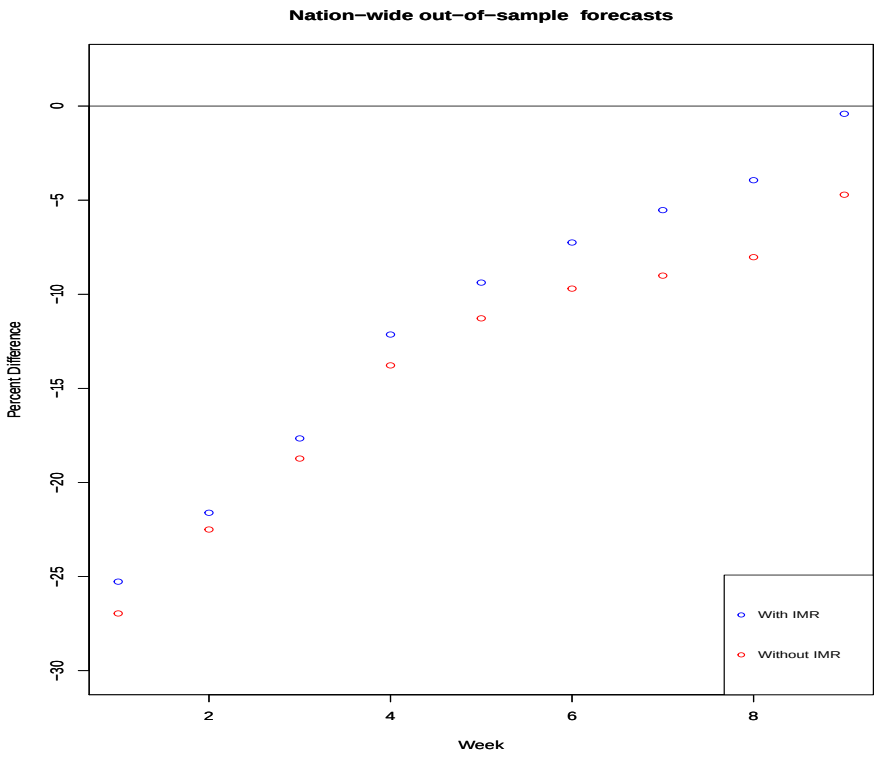
	week $t+1$	IMR	WithoutIMR
1	Apr-20-2020	-25.27	-26.96
2	Apr-27-2020	-21.61	-22.50
3	May-4-2020	-17.66	-18.73
4	May-11-2020	-12.14	-13.78
5	May-18-2020	-9.38	-11.28
6	May-25-2020	-7.25	-9.70
7	June-1-2020	-5.53	-9.01
8	June-8-2020	-3.94	-8.03
9	June-15-2020	0.41	-4.71

the information up-to the previous week. Out-of-sample forecast errors are considered more relevant for policy than in-sample forecast errors. Thus, we have a strong result favoring the use of IMR-based bias corrections in predicting nationwide Covid-19 new deaths.

Figure 19 also reveals that the IMR value is decreasing over time for each state, which is indicative of improved testing administration. An exciting policy implication of our study of the bias measured by IMR is that public health strategies can target resources to 'states' with high IMR values to help control the spread of infection.

Table V has a correlation matrix with certain abbreviated row and column titles. In defining the titles, we let "Z%Δ" denote a percent difference for some "Z" defined below averaged over nine weekly values for each state starting with the first week starting on April 20, 2020, and ending in the ninth week starting on June 15, 2020. When Z="Fitted" in the title, we have a set of fifty-one %Δ between the fitted values of the Poisson model with and without the IMR. When Z="IMR," the model has the bias index IMR included as a regressor, and the percent difference is calculated between the fitted values and the actual (out-of-sample) cumulative deaths for the next week. When Z="NoIMR," the model does not have IMR, and

Figure 19: Scatter plot of out-of-sample nation-wide death forecasts by week



we calculate the percent difference between the fitted values and the actual (out-of-sample) cumulative deaths for the next week. The title "Democrat%" refers to the percent Democratic vote in the last general election. We denote statistically insignificant correlation coefficients (p-values exceeding 0.05) by the superscript 'n.'

Table V suggests that higher Democratic voting states have mostly lower percent errors except in models without the IMR as a regressor, where the insignificant positive coefficient 0.1278 has a large p-value of 0.37. The correlation between IMR%Δ and NoIMR%Δ is insignificantly negative –0.1307, with a large p-value of 0.36. Near-zero correlation means that with and without IMR values are not related to each other. That is, IMR provides independent information, not already available. Is the independent informa-

TABLE V

CORRELATION MATRIX BETWEEN THREE CHOICES OF Z IN Z%Δ AND PERCENT OF DEMOCRATIC VOTES IN THE LAST ELECTION IN THAT STATE.

	Fitted%Δ	IMR%Δ	NoIMR%Δ	Democrat%
Fitted%Δ	1.0000	0.9256	-0.4628	-0.4249
IMR%Δ	0.9256	1.0000	-0.1307^n	-0.4498
NoIMR%Δ	-0.4628	-0.1307^n	1.0000	0.1278^n
Democrat%	-0.4249	-0.4498	0.1278^n	1.0000

tion useful? Yes, because Table IV shows that the bias correction by IMR improves out-of-sample forecasts of new deaths from Covid-19.

6. STATEWISE FORECASTS OF INFECTIONS

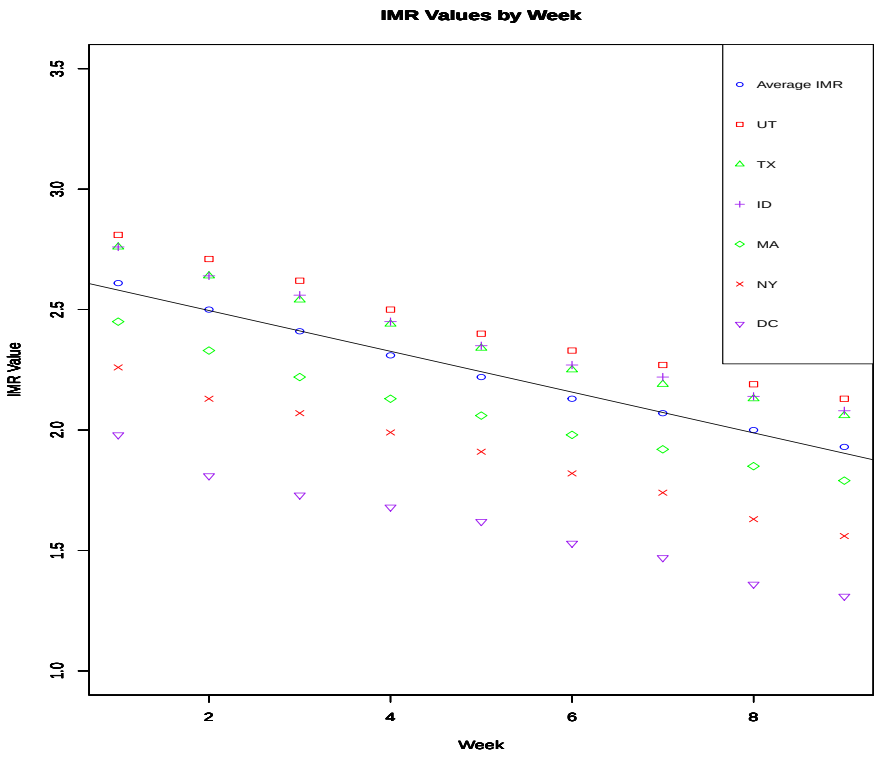
This study focused on the estimation of deaths as a result of cumulative infections. We know from Section 2 that there is a bias associated with reported infection count due to non-random sampling (i.e., gaps in testing administration and/or efficiency). We adjust for the bias by applying the inverse Mills ratio to the Poisson model for death count. It is possible to study infections directly by re-estimating the models with new infections rather than new deaths as the dependent variable.

Instead of repeating all calculations, we claim that some information about new infections is already available from our estimates of IMR for a model with new deaths. Our estimated IMR values can serve as approximate indicators of biases in estimated infections also. More specifically, we suspect that low inverse Mills ratio values indicate (relatively) more reliable forecasts of disease spread, while higher inverse Mills ratio values suggest (relatively) greater testing bias. Furthermore, we predict that 'states' with higher inverse Mills ratio values may see unanticipated spikes in infection count (reported and non-reported) during the upcoming weeks.

The 'states' with the highest average inverse Mills ratio values across all weeks are (UT, ID, TX). Interestingly, these are among the states with suddenly high infections reported on June 26. The 'states' with the lowest

average IMR values across all weeks are (MA, NY, DC) are not reporting sudden increases in infections on June 26. Figure 20 plots the average inverse Mills ratio value for each week compared to the inverse Mills ratio values for the selected 6 'states' (UT, ID, TX, MA, NY, DC) over 9 weeks. The overall downward slopes for all states suggest that the bias is decreasing over time.

Figure 20: Plots IMR Values by Week from 2020/04/20 to 2020/06/15 for the 'states' with the highest and lowest values.



7. FINAL REMARKS

We find that econometric tools, including the inverse Mills ratio (IMR) to assess the bias in data, are successful in achieving superior forecasts of deaths related to Covid-19. In particular, we believe high IMR values indi-

cate (relatively) larger testing bias due to non-random sampling, inefficiency, and gaps in test administration between specific regional populations.

A negative correlation between IMR and slope of ρ the predictive coefficient suggests that as IMR increases, the slope decreases within various geographical groups of states further divided into democrat and republican affiliations of their governors who have a crucial role in fighting Covid-19 pandemic.

This study has established that we can improve the forecasts of deaths with the use of IMR for bias correction. These tools, mostly ignored by epidemiologists, have several potential applications in all kinds of epidemics where the disease testing is biased for whatever reason. It is not possible to describe the full range of potential applications of IMR methods to remove data biases in Covid-19 studies around the world. One could easily focus on alternative outcomes of interest, such as predicting the demand for hospital beds, particular devices (e.g., ventilators) infections instead of deaths by changing the dependent variable in our outcome equations.

We report a focused 11-week study of trends in new deaths predicted by lagged cumulative infections based on a simplified model without bias correction. It finds that the desired negative direction in these trends is not statistically significant, suggesting that the direction can reverse in the future. We have not computed the statistical significance of trends from the quadratic specification of trends. If we ignore statistical significance, we have mostly negative quadratic trends in 42 states. Unfortunately, seven of nine states with positive quadratic trends have Republican governors. Since statistical significance of quadratic model results is not known, these results do not necessarily mean that Republican governors need to do more to control Covid-19.

The model can be applied to more aggregate (international) or disaggregate data. For example, a similar analysis may be conducted within indi-

vidual 'states' using county (census tract) level data. Also, these methods can assess the impact of specific relaxation policies and predict underestimation by focusing on coefficients of specific explanatory variables. For example, one can utilize the model described in this paper to correctly assess the effect of poverty on the burden of Covid-19 while using the IMR to remove potential underestimation in computing the impact of poverty on the disease outcome.

It would be interesting to use the R package for maximum entropy bootstrap (meboot) to construct 95% confidence intervals around forecast errors. Fenga (2020) has already used the 7-step meboot algorithm described in Vinod and López-de-Lacalle (2009) to study Covid-19 spread in Italian provinces.

Since IMR bias correction has improved total new death forecasts in nine of nine weeks, one can be confident that bias correction is working for this problem. It would be possible to report weekly state-wise forecasts of bias-corrected new deaths until the pandemic ends.

REFERENCES

Fenga, L., 2020. Forecasting the CoViD-19 diffusion in Italy and the related occupancy of Intensive Care Units. Technical Report. Italian National Institute of Statistics ISTAT, Rome, Italy 00184. URL: https://www.medrxiv.org/content/10.1101/2020.03.30.20047894v1.full.pdf.

Heckman, J.J., 1979. Sample selection bias as a specification error. Econometrica 47, 153–161.

Kermack, W.O., McKendrick, A.G., 1991. Contributions to the mathematical theory of epidemics-i. Bulletin of Mathematical Biology 53(1-2), 33–55. URL: https://doi.org/10.1007/BF02464423.

Toomet, O., Henningsen, A., 2008. Sample selection models in R: Package sampleSelection. Journal of Statistical Software 27. URL: http://www.jstatsoft.org/v27/i07/.

Vinod, H.D., 2008. Hands-on Intermediate Econometrics Using R: Templates for Extending Dozens of Practical Examples. World Scientific, Hackensack, NJ. URL:

https://www.worldscientific.com/worldscibooks/10.1142/6895. ISBN 10-981-281-885-5.

Vinod, H.D., López-de-Lacalle, J., 2009. Maximum entropy bootstrap for time series: The meboot R package. Journal of Statistical Software 29, 1–19. URL: http://www.jstatsoft.org/v29/i05/.