EIEF WORKING PAPER SERIES

EINAUDI INSTITUTE FOR ECONOMICS AND FINANCE

EIEF Working Paper 21/18

December 2021

A note of caution in interpreting cross-country correlations of COVID-19 vaccination and infection rates

By

Francesco Bartolucci
(University of Perugia)

Franco Peracchi
(University of Rome Tor Vergata and EIEF)

Daniele Terlizzese
(EIEF)

A note of caution in interpreting cross-country correlations of COVID-19 vaccination and infection rates*

Francesco Bartolucci
University of Perugia, Italy

Franco Peracchi
University of Rome Tor Vergata and EIEF, Italy

Daniele Terlizzese
EIEF, Italy

December 10, 2021

Abstract

We show that the finding in Subramanian and Kumar (2021) of no discernible linear relationship between COVID-19 vaccination rates and new infection rates in a cross-section of countries is misleading, because it ignores the substantial degree of heterogeneity across countries. The latter reflects large differences in the stages of the infection and the vaccination process, healthcare systems, people's attitude towards vaccination, etc. In the presence of such heterogeneity simple correlations are hardly interpretable. This is a well-known phenomenon, sometimes referred to as the Simpon's paradox. Exploiting longitudinal data, they show that the estimated linear relationship becomes negative and statistically significant when controlling for time-invariant differences across countries.

Keywords: COVID-19 vaccination rates, COVID-19 infection rates, cross-country heterogeneity, Simpson's paradox.

*Corresponding author: Franco Peracchi, University of Rome Tor Vergata, Rome, Italy (franco.peracchi@uniroma2.it).

Introduction

One of the side effects of the virulent diffusion of COVID-19 has been the increasing familiarity of the general public with statistical concepts. We daily peruse levels, percentages, averages, and rates of change of a phenomenon that is affecting our life in dramatic and unexpected ways. Yet, we did not became trained statisticians in one day, and it is all too easy to fall prey of pitfalls in reading and interpreting the data that we see on our screens.

Students of Statistics 101 usually learn that correlation does not imply causation, a lesson that sadly eludes many journalists, pundits, and politicians that we hear commenting on the evolution of the pandemic.

Even correlations, in and of themselves, can however be misleading. This is particularly true when looking at the correlation between the diffusion of the infection, which is changing over time and at different speed in different places, and the vaccination campaign, which is conducted in different countries with different delays and whose effects take time to materialize and are not constant in time. In these cases, cross-country correlations, without controlling for the different stages of the two phenomena in different countries, are hardly interpretable. Unfortunately, if these correlations are offered to the voracious appetites of the media, they end up being interpreted in ways that are inevitably wrong, and often colored by ideological bents.

An instance of a correlation that is not interpretable is provided by a recent communication by S.V. Subramanian and A. Kumar [1] (henceforth SK), that looks at the role of vaccination rates as the primary mitigation strategy to combat COVID-19 around the world and finds no evidence of a negative correlation, either contemporaneous or lagged one month, between the percentage of population fully vaccinated and new COVID-19 cases across both 68 countries and 2,947 U.S. counties. Based on this evidence, SK conclude that other non-pharmaceutical prevention efforts "needs to be renewed in order to strike the balance of learning to live with COVID-19 in the same manner we continue to live a 100 years later with various seasonal alterations of the 1918 Influenza virus."

Although we do not disagree with these conclusions, it would be wrong to interpret the findings in SK as evidence that vaccination rates do not help reduce COVID-19 infection. Rather, they represent a neat illustration of a well-known pitfall, sometimes referred to as the Simpson's paradox [2], which arises from the substantial degree of cross-country heterogeneity in the stages of the infection and the vaccination processes, as well as in healthcare systems, people's attitude towards vaccination, etc. Interpreting a zero or positive cross-country correlation as evidence of no effect of vaccination rates on infection, besides having no sound statistical basis, contributes to distort the public policy debate.

Methods

As SK, we focus on the correlation between the COVID-19 infection rate (henceforth denoted by Y), defined as the total number of new COVID-19 cases per million people, and the vaccination rate (henceforth denoted by X), defined as the percentage of the population that is fully vaccinated against COVID-19. We employ the same data and sample selection criteria used by SK. In particular, we use the data provided by Our World in Data [3] for the set of 68 countries considered by SK. This is done for comparability with SK and does not represent an endorsement of their sample selection criteria. Indeed, there are several quirks in the sample selection: most European countries are missing, while some very small and unrepresentative countries are included.

Unlike SK, who only use the cross-sectional information available for a specific reference date, namely September 3, 2021, we take advantage of the full longitudinal information on daily vaccination and infection rates starting from the beginning of the vaccination campaign in January 2021 through September 3, 2021, the SK reference date. This allows us to control for time-invariant differences across countries, in order to take into account, albeit imperfectly, the cross-country heterogeneity that makes the contemporaneous correlation uninterpretable. Further, to limit the impact of inaccurate or missing daily information, we switch from daily data to weekly averages for X and weekly sums for Y, with weeks defined as the seven-day periods from Monday to Sunday.

We compare the results obtained from a simple ordinary least squares (OLS) regression of Y on X, using either the cross-sectional information for the reference week of September 3, 2021 (this is the correlation reported in SK, corresponding to the "trend line" in their Fig. 1) or the information from all 35 weeks of 2021 up to that week, with those obtained using the longitudinal nature of the data and two standard ways of controlling for time-invariant differences across countries: a "first differences" (FD) model that regresses weekly changes in Y (denoted by ΔY) on weekly changes in X (denoted by ΔX) and a "fixed effects" (FE) model that regresses Y on X and a set of binary country indicators. All regressions are unweighted, as in SK.

Findings

The first column of Table 1 presents the OLS estimates of the intercept and slope of the simple linear regression of Y on X using only the cross-sectional information for the SK reference week of September 3, 2021, while the second column presents the estimates from all 35 weeks of 2021 up to that week. The remaining two columns of the table show that the estimated

slope of the linear relation linking Y to X switches from positive to negative when we exploit the longitudinal nature of the data using either the FD or the FE model. This sharp change reflects the bias in the cross-sectional OLS estimates, which arises from ignoring time-invariant differences across countries. The third column presents the estimates of the FD model using information only for the SK reference week and its previous week. It reveals that even minimal longitudinal information is enough to flip the sign of the estimated regression slope. Statistical significance is low, however, because of the noise in the differenced data and the limited sample size. The last column presents the results of the FE model that exploits the full longitudinal information available. Now the slope coefficient is not only negative but also strongly statistically significant because of the larger sample size.

Figure 1 shows the scatterplot of Y and X for the SK reference week (the hollow red dots), the positively sloped OLS lines fitted to this scatterplot (in red), the scatterplot of Y and X for all weeks of 2021 up the reference week (the hollow grey dots), and the negatively-sloped FE line fitted to this scatterplot (in blue).

Qualitatively, results do not change when we regress Y on X lagged 4 weeks, or we stick to daily data (as in SK), or we control for other regressors (e.g., the stringency index, indicators of the population age structure, etc.), or we allow for nonlinearities in the relationship between Y and X, or we consider alternative vaccination measures that weight differently the fraction of people who received one, two, or three doses. In all these cases, the relation linking Y and X remains negative and statistically significant.

Interpretation

Our findings are an illustration of the Simpson's paradox [2], a term used to denote the case in which a trend that appears when combining a number of heterogeneous groups (countries in our case) disappears or reverses when we control for group heterogeneity.

There are many reasons for cross-country heterogeneity in the relationship between COVID-19 infection and vaccination rates: different stages of the infection and the vaccination process, differences in healthcare systems, differences in people's attitude towards vaccination, etc.

Interpreting the positive or null contemporaneous cross-country correlation as evidence of the absence of a negative effect of vaccination on COVID-19 infection rates is therefore wrong, and contributes to distort the public policy debate, as happened in recent weeks in many countries, including our own. Researchers with a sound statistical training and scientific journals should know better than fanning, with hardly interpretable result, the embers of vaccine resistance.

References

1. Subramanian SV, Kumar A. Increases in COVID-19 are unrelated to levels of vaccination across 68 countries and 2947 counties in the United States. European Journal of Epidemiology. 2021. https://doi.org/10.1007/s10654-021-00808-7.

2. Simpson EH. The interpretation of interaction in contingency tables. Journal of the Royal Statistical Society–Series B. 1951. http://www.jstor.org/stable/2984065.

3. Ritchie H, Ortiz-Ospina E, Beltekian D, Mathieu E, Hasell J, Macdonald B, Giattino C, Appel C, Rodés-Guirao L, Roser M. Coronavirus pandemic (COVID-19). 2020. Published online at OurWorldInData.org. Retrieved from: https://ourworldindata.org/coronavirus.

Variable	Model 1	Model 2	Model 3	Model 4
X	17.436	11.982		-9.663
	(7.868)	(1.656)		(1.713)
	[2.216]	[7.234]		[-5.640]
ΔX			-14.811	
			(37.523)	
			[-.395]	
Intercept	699.731	856.940	1.655	1078.380
	(311.125)	(33.077)	(86.061)	(28.715)
	[2.249]	[25.907]	[.019]	[37.555]
N	68	2380	68	2380
R^2	.0693	.0215	.0024	.0136

Table 1: Model estimates. Model 1 is a simple linear regression of Y on X estimated by OLS on the reference week of September 3, 2021, Model 2 is the same model estimated from all 35 weeks of 2021 up to that week, Model 3 is the FD model estimated on the reference and previous week, and Model 4 is the FE model estimated on all 35 weeks of 2021 up to the reference week. Standard errors in parentheses, t-ratios in square brackets.

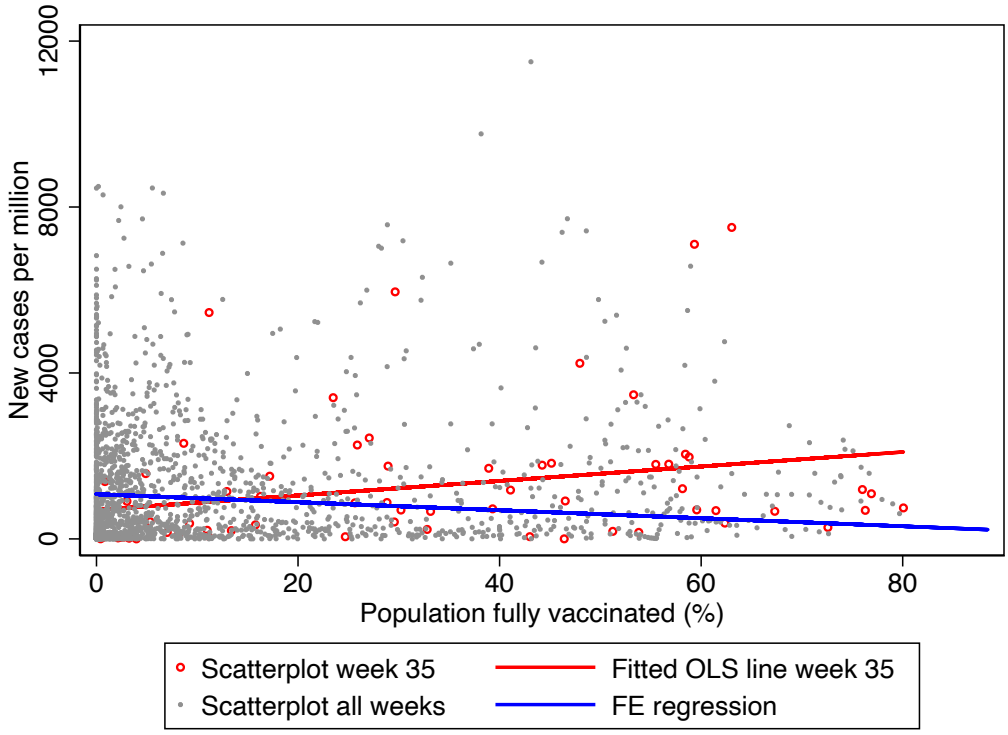

Figure 1: Scatterplot of Y and X for the reference week of September 3, 2021 (hollow red dots), OLS lines fitted to this scatterplot (in red), scatterplot of Y and X for all weeks of 2021 up the reference week (hollow grey dots), and FE line fitted to this scatterplot (in blue).