scientific reports

Check for updates

OPEN

MATTERS ARISING

Pairwise difference regressions are just weighted averages

Carlos Góes

ARISING FROM: R. F. Savaris et al.; *Scientific Reports* https://doi.org/10.1038/s41598-021-84092-1 (2021).

Savaris et al.[1] aim at "verifying if staying at home had an impact on mortality rates." This short note shows that the methodology they have applied in their paper does not allow them to do so. An estimated coefficient $\beta \approx 0$ does not imply that there is no association between the variables in either country. Rather, their pairwise difference regressions are computing coefficients that are weighted-averages of region-specific time series regressions, such that **it is possible that the association is significant in both regions but their weighted-average is close to zero**. Therefore, the results do not back up the conclusions of the paper.

Consider two regions: A and B. Suppose that the true relationships between the change in deaths per million (ΔY_t^i) and the change in an index of staying at home (ΔX_t^i) at epidemiological week t in countries $i = A, B$ are the following:

$$\begin{aligned}\Delta Y_t^A &= \beta_A \Delta X_t^A + \varepsilon_t^A \\ \Delta Y_t^B &= \beta_B \Delta X_t^B + \varepsilon_t^B\end{aligned}$$

For simplicity in exposition, assume that $\Delta X_t^A, \Delta X_t^B, \varepsilon_t^A, \varepsilon_t^B$ are all zero mean, iid processes. By subtracting the second equation from the first and defining $\Delta Y_t \equiv \Delta Y_t^A - \Delta Y_t^B$ and $\Delta X_t \equiv \Delta X_t^A - \Delta X_t^B$, we can write:

$$\begin{aligned}\Delta Y_t^A - \Delta Y_t^B &= \beta(\Delta X_t^A - \Delta X_t^B) + (\beta_A - \beta)\Delta X_t^A - (\beta_B - \beta)\Delta X_t^B + (\varepsilon_t^A - \varepsilon_t^B) \\ \Delta Y_t &= \beta \Delta X_t + \eta_t\end{aligned} \tag{1}$$

where $\eta_t \equiv (\beta_A - \beta)\Delta X_t^A - (\beta_B - \beta)\Delta X_t^B + (\varepsilon_t^A - \varepsilon_t^B)$. It is easy to see that, for $\beta_i \neq \beta$, estimation of β will not be consistent, since, by construction, $cov(\Delta X_t, \eta_t) \neq 0$.

If nonetheless one estimates (1) by ordinary least squares, what does the regression coefficient β converge to? It turns out that it converges to a variance-weighted average of β_A, β_B, as summarized in the following proposition.

Proposition 1 *Let $\Delta X_t^A, \Delta X_t^B, \varepsilon_t^A, \varepsilon_t^B, \beta_A, \beta_B, \beta$ be all as above. Then $\hat{\beta}$, the ordinary least squares coefficient of regressing ΔY_t on ΔX_t, converges in probability to*:

$$\beta = w\beta_A + (1 - w)\beta_B \tag{2}$$

with $w \equiv \frac{\mathbb{E}[(\Delta X_t^A)^2]}{\mathbb{E}[(\Delta X_t^A)^2] + \mathbb{E}[(\Delta X_t^B)^2]}$.

Proof Under the stated assumptions, $\hat{\beta} = \frac{\sum_t^T \Delta X_t \Delta Y_t}{\sum_t^T \Delta X_t^2} \xrightarrow{p} \frac{\mathbb{E}[\Delta Y_t \Delta X_t]}{\mathbb{E}[\Delta X_t^2]} \equiv \beta$. One can calculate the population parameter β analytically:

$$\begin{aligned}\beta &= \frac{\mathbb{E}[\Delta Y_t \Delta X_t]}{\mathbb{E}[\Delta X_t^2]} \\ &= \frac{\mathbb{E}[(\Delta Y_t^A - \Delta Y_t^B)(\Delta X_t^A - \Delta X_t^B)]}{\mathbb{E}[(\Delta X_t^A - \Delta X_t^B)^2]} \\ &= \frac{\mathbb{E}[\Delta Y_t^A \Delta X_t^A] + \mathbb{E}[\Delta Y_t^B \Delta X_t^B]}{\mathbb{E}[(\Delta X_t^A)^2] + \mathbb{E}[(\Delta X_t^B)^2]} \qquad \because \mathbb{E}[\Delta X_t^A \Delta X_t^B] = \mathbb{E}[\Delta X_t^A \Delta Y_t^B] = \mathbb{E}[\Delta X_t^B \Delta Y_t^A] = 0 \\ &= \frac{\mathbb{E}[(\Delta X_t^A)^2]}{\mathbb{E}[(\Delta X_t^A)^2] + \mathbb{E}[(\Delta X_t^B)^2]} \frac{\mathbb{E}[\Delta Y_t^A \Delta X_t^A]}{\mathbb{E}[(\Delta X_t^A)^2]} + \frac{\mathbb{E}[(\Delta X_t^B)^2]}{\mathbb{E}[(\Delta X_t^A)^2] + \mathbb{E}[(\Delta X_t^B)^2]} \frac{\mathbb{E}[\Delta Y_t^B \Delta X_t^B]}{\mathbb{E}[(\Delta X_t^B)^2]}\end{aligned}$$

Department of Economics, University of California, San Diego, USA. email: cgoes@ucsd.edu

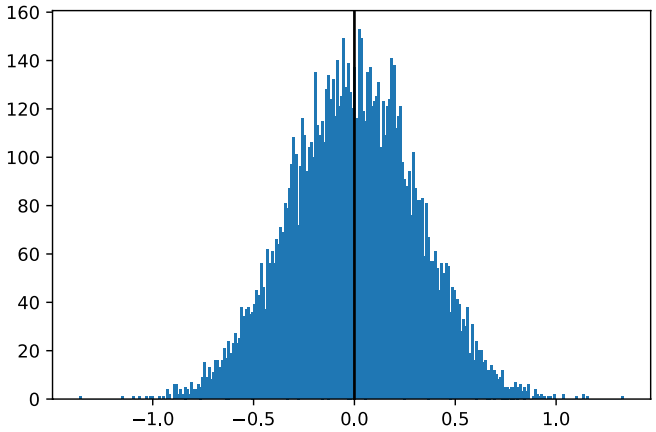

Figure 1. In-sample simulated $\hat{\beta}$ for 10,000 random draws with $\Delta X_t^i \sim N(0, 10)$, $\varepsilon_t^i \sim N(0, 1)$, and $\Delta Y_t^i = \beta_i \Delta X_t^i + \varepsilon_t^i$, for $i = A, B$; $T = 1,000$; and $\beta_A = 10$, $\beta_B = -10$. As expected the sample values are distributed around the true population value of $\beta = 0$.

Note that $\frac{\mathbb{E}[\Delta Y_t^A \Delta X_t^A]}{\mathbb{E}[(\Delta X_t^A)^2]} = \beta_A$ and $\frac{\mathbb{E}[\Delta Y_t^B \Delta X_t^B]}{\mathbb{E}[(\Delta X_t^B)^2]} = \beta_B$. Using that and the definition of w we arrive at the desired result. □

The intuition regarding the (2) in the Proposition is simple. Whenever the variance of ΔX_t^A is large relative to country B, $w \to 1$ and $\beta \to \beta_A$. Similarly, if the variance of ΔX_t^B is large relative to country A, $w \to 0$ and $\beta \to \beta_B$.

What does this mean for the analysis of Savaris *et al.*[1]? In general, it means that one cannot interpret their estimated $\hat{\beta}$ without knowing the underlying relative variances. Additionally, one cannot infer that an insignificant (or even numerically zero) $\hat{\beta}$ implies absence of association in either country.

To see that, suppose countries A and B have identical variance in their independent variables, but β_A, β_B are different. In country A, the policymaker adjusts stay-at-home orders in response to the increase in deaths, such that the change in the percentage of the public staying at home is positively correlated with the change in deaths. In country B, the policymaker does not act, such that the change in share of population staying at home is negatively correlated with contacts, infections, and deaths.

Consider the case in which $\beta_B = -\beta_A$. Then, since the regions have identical variance, $w = 1/2$ and $\beta = 0$ even though the true association is nonzero in both countries. The regression coefficients in Savaris *et al.*[1] should not lead one to conclude that, in either country, there is no association between the change in mobility and the change in deaths per million. Figure 1 shows the result of 10,000 simulated $\hat{\beta}$ in which $\beta_A = 10$ and $\beta_B = -10$. In this case, $var(X_t^A) = var(X_t^B)$ and variables are iid and normally distributed. As expected, sample estimates are distributed around the population value of $\beta = 0$.

For $\beta_A \neq \beta_B$, then, region-specific dynamics are heterogeneous and, as shown by Pesaran & Smith[2], aggregating or pooling slopes can lead to biased estimates, making individual regressions for each group member preferable. If authors assume that $\beta_A = \beta_B$ for each pair in their sample – i.e., homogeneous β –, then dynamic panels would have many advantages in terms of efficiency and use of instruments to circumvent endogeneity. In either case, their pairwise approach would not be appropriate.

In order to verify if "staying at home had an impact on mortality rates," it would be necessary to address many other issues in the analysis, including, but not limited to, omitted variable bias, measurement error, and endogeneity of the regressors. However, as shown above, even in a purely correlational analysis, with no causality claims, the applied methodology will simply deliver a weighted-average of coefficients across the two regions. An estimated coefficient $\beta \approx 0$ does not imply that there is no association between the variables in either country. Therefore, their conclusion does not follow from their regressions.

Received: 19 March 2021; Accepted: 10 November 2021
Published online: 29 November 2021

References

1. Savaris, R. F., Pumi, G., Dalzochio, J. & Kunst, R. Stay-at-home policy is a case of exception fallacy: an internet-based ecological study. Scientific Reports 11, 5313. issn: 2045-2322 (2021).
2. Pesaran, M. H. & Smith, R. Estimating long-run relationships from dynamic heterogeneous panels. *J. Econ.* **68**, 79–113 (1995).

Author contributions

This article is solo authored.

Competing interests
The author declares no competing interests.

Additional information
Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-021-02096-3.

Correspondence and requests for materials should be addressed to C.G.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.