Editorial

Matching in Difference-in-Differences: between a Rock and a Hard Place

There is intense interest in methods to measure the impacts of nonrandomized policy interventions. Difference-in-differences analysis is one such method that is increasingly popular in health policy evaluations. In 2015, Ryan, Burgess, and Dimick published a manuscript in *HSR* on the assumptions and optimal specifications for difference-in-differences (Ryan, Burgess, and Dimick 2015). We have referred to this paper and are grateful for Andrew Ryan's commentary on our manuscript, published in this issue of *HSR*, which explores how difference-in-differences using matched samples may be vulnerable to bias due to regression to the mean (RTM).

In response to his comments, we argue that the situations susceptible to RTM bias are common and that the benefits of matching on trend may be small compared to the confidence this approach engenders. In addition, we illustrate how RTM bias affects the simulations both in Ryan et al. (2015) and in our paper.

Our paper demonstrates the statistical phenomenon of RTM in difference-in-differences analyses that match on time-varying variables in the preperiod. As theory predicts, when matching results in selection of units with extreme measurements of a time-varying variable at one point in time, subsequent measurements taken on those matched units will be closer to their population mean(s). The intuition for how matching can bias difference-in-difference studies is simple: Researchers enforce a constraint by matching in the preperiod and then release that constraint in the postperiod, creating a spurious effect as the unconstrained units "bounce back." The magnitude of RTM will depend on how extreme the selection is (i.e., how different the groups are at baseline), and how unstable the matching variables are over time (i.e., serial correlation).

In his commentary, Ryan argues that only a narrow set of scenarios would cause matching to produce biased effects due to RTM. Indeed, our simulations demonstrate large bias only when the matching variable has modest serial correlation and the difference in level between the two groups is large and stable. However, we disagree that these situations are unlikely to be encountered in practice. We have identified many cases in the literature of matching on time-varying variables, including preperiod outcomes. Further, one of the most challenging aspects of this phenomenon is that the greater the difference in level between the treatment and control groups, the greater the potential rationale for matching, and the greater the potential bias due to regression to the mean. In other words, a researcher's temptation to match and the risk of RTM bias are highly correlated.

Ryan notes that researchers usually implement matching in response to evidence that the assumptions of difference-in-differences do not hold, that is, when preperiod differences in outcome trend are observed. We observe that researchers match in response to preperiod differences in both levels and trends of outcomes and covariates, often appealing to a general desire "to improve comparability between the treatment and control group." Researchers may even prespecify that matching will be implemented prior to inspecting these differences.

In fact, in their 2015 paper, Ryan and colleagues offer matching as a solution to extremely high false-positive rates in a simulation with no preperiod trend differences, only level differences. Their Table 2 shows false rejection rates of 98–100 percent in the unmatched cases that drop to 1–1.5 percent with matching. This is despite a scenario that apparently satisfies the DID assumptions: Treatment assignment is only related to level, not trend. In a small simulation, described below, we demonstrate that RTM bias underlies both their results and ours.

For clarity, our simulations only included one variable in the matching algorithm. In practice, even if matching is only applied in response to evidence of a differential preperiod trend, the common "kitchen sink" approach results in many matching variables and may include time-varying variables that are prone to mean reversion. The aggregate bias introduced by regression to the mean in such multivariate matching scenarios is difficult to determine a priori.

Address correspondence to Jamie R. Daw, Ph.D., Department of Health Care Policy, Harvard Medical School, 180 Longwood Avenue, Boston, MA 02115; e-mail: jdaw@fas.harvard.edu. Laura A. Hatfield, Ph.D., is with the Department of Health Care Policy, Harvard Medical School, Boston, MA.

Ryan also suggests that our results may not be generalizable to many real-world cases because in our simulations “there is no overlap in the true distributions” of the matching variables in the treatment and comparison groups. This is simply not accurate. Our simulation imposes a difference in the population means of the matching variable (outcome or covariate) between the treatment and control groups; however, the population distributions overlap in all scenarios. We believe that what Ryan means is that the two groups are not sampled from the *same* distribution.

In their 2015 paper, Ryan and colleagues made the probability of treatment increase with the preperiod mean outcome, so the treated and control groups were biased samples from the *same* population. That is why they observed such high false-positive rates in the unmatched analysis (their Table 2). The two groups had different means because they were separated by the treatment assignment mechanism; that was the “constraint” in the preperiod. In the postperiod, the two groups regressed back to the same mean from different places. Contrast this to our scenario in which samples from two different populations were constrained (by matching) to be *similar* in the preperiod. In the postperiod, the two groups regressed back to different means from the same place. Perniciously, matching is both the *solution* to RTM bias in the Ryan et al. simulation and the *cause* of RTM bias in our simulation.

If we observe preperiod mean differences in level between the treated and control groups, we could be living in either of these two worlds. In the first world, treated and control units come from the same population (with the same mean) and treatment assignment is correlated with preperiod level, creating a preperiod level difference that will disappear in the postperiod. Here, the unmatched analysis is biased due to RTM and matching on preperiod level reduces the bias. In the second world, treated and control units come from different populations with different means. Here, the unmatched analysis is unbiased because the difference between the groups is stable, but matching on preperiod level introduces RTM bias.

Figure 1 illustrates how similar data from these two worlds appear in the preperiod and the opposite effects of matching in each. In panels (A) and (B), we show data on 2,000 units simulated from a normal distribution with mean 0 in each of six periods. We imagine a null intervention divides the periods into 3 pre and 3 post. The probability of each unit being assigned to treatment increases over quartiles of the preperiod means of each unit as follows: 0.35, 0.45, 0.55, 0.65. That is, units with higher-than-average means in the preperiod are more likely to be assigned to treatment, and those with lower-than-average means are more likely to be assigned to control. The result is

Figure 1: Simulation Results

Notes. Simulated data for treated (black) and control (gray) units that come from the same population (top row) or different populations (bottom row) and are either matched on preperiod means of the outcome (right column) or unmatched (left column). Dashed vertical line marks the start of the postperiod.

approximately 1,000 treated and 1,000 control units. This is a simplification of the Ryan et al. simulation scenario and illustrates the separation in level that this treatment assignment mechanism induces: The mean is 0.11 in the treated group and -0.19 in the control group, even though all the observations were simulated from a normal distribution with mean 0. However, as panel (A) shows, this difference disappears in the postperiod due to RTM. Panel (B) shows how matching solves this by constraining the two groups to be similar in the preperiod (as they will be in the postperiod).

In panels (C) and (D), we show data on 1,000 treated units simulated from a normal with mean 0.11 and 1,000 control units simulated from a normal with mean -0.19 . The preperiod mean difference is therefore the same as in the Ryan et al. scenario. The two worlds are indistinguishable in the preperiod: Compare panels (A) and (C) prior to the “intervention,” marked by the

vertical dashed line. In panel (C), we see that the level difference persists over the whole study period because the mean difference between the two groups is stable. Panel (D) shows the RTM bias induced by matching in this case; when the matching constraint is lifted, the groups regress back to their respective means.

How can we tell whether observed preperiod level differences reflect transient differences (as when treatment assignment is determined by recent preperiod outcome means) or stable differences (as when treated and control units come from different populations)? As Figure 1 shows, preperiod observations generated by these two scenarios will look identical, making differentiation difficult or impossible by empirical checks alone. Instead, researchers must carefully think through the possible treatment assignment mechanisms that may be operating in the real-life situation they are investigating. For example, if researchers are aware that a pay-for-performance incentive was assigned to physicians within a state based on average per-patient spending in the past year, one may be comfortable assuming that treatment assignment mechanism is operating at the unit level (i.e., the potential treatment and control units are from the same population). In contrast, if the same incentive was assigned to all physicians within a state and a researcher chooses a control state based on geographic proximity, it may be more reasonable to assume that treatment assignment is operating at the population level (i.e., the potential treatment and control units are from separate populations).

Ryan also notes that we find matching on preperiod outcome trend reduces bias in scenarios with high serial correlation of the outcome. We argue that this bias reduction is trivial compared to the overconfidence it can produce. Matching on preperiod trend produces a matched sample with similar trends in the treatment and control groups *by construction*. This is usually interpreted as strong evidence that the parallel trends assumption holds in the matched sample. However, this confidence may be unwarranted. The actual benefit depends on the stability of trends in the outcome over time. As we show in our simulation, when trends are not stable, the units tend to revert to their population mean trend in the postperiod. While this mean reversion will not result in *additional* bias relative to unmatched samples, when trends are not highly stable over time, it will result in trivial or no reduction in bias coupled with a conviction that the assumptions of the analysis hold. We would argue that the hidden bias that results from matching on transient, random trends is especially dangerous.

So when might matching help? We agree with Ryan that matching has the potential to be beneficial in difference-in-differences if the matching

variables are correlated with future outcomes. In addition, if we have strong theoretical or subject matter evidence that treatment and control groups come from the same population, matching can reduce RTM bias as the Ryan et al. results demonstrate. Fixed variables (not subject to RTM) that are strongly predictive of future outcome trends are the “holy grail” of matching variables in difference-in-differences. However, while we can use theory and empirical evidence, we can rarely identify observed variables that predict future trends (and thus meaningfully reduce bias). In fact, identifying variables that reliably predict future health care spending trends could make us all very rich.

Are techniques like synthetic control groups the answer? Depending on the variables used to construct the synthetic control weights, they may be subject to the same concerns. The synthetic control method has a strong parallel to matched difference-in-differences. In this method, the control group is a weighted combination of potential control units with weights chosen based on similarity to the treated units on preperiod characteristics. The preperiod measures used in synthetic control construction include preperiod outcome measurements. Like matching, weighting is a form of asymmetric selection of units based on preperiod measurements of time-varying variables and treatment effect estimation based on the difference between the two groups in the post-period. Based on theory, there is no reason to believe that this preperiod weighting exercise would be immune to RTM effects.

In summary, we suggest that researchers remain alert to the threats of RTM bias in both the direction identified by Ryan and colleagues and the direction we have identified. Investigators must thoughtfully consider likely treatment assignment mechanisms, potential time-varying confounders, and the temporal correlation structure of their outcomes of interest before proceeding with weighting or matching on any time-varying variables.

ACKNOWLEDGMENTS

Joint Acknowledgment/Disclosure Statement: The authors have no conflict of interests to declare.

Disclosures: None.

Disclaimer: None.

Jamie R. Daw
Laura A. Hatfield

REFERENCES

- Ryan, A. M., J. F. Burgess, and J. B. Dimick. 2015. "Why We Should Not Be Indifferent to Specification Choices for Difference-in-Differences." *Health Services Research* 50 (4): 1211–35.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

Appendix SA1: Author Matrix.