Claremont Graduate University

ECON 383: Causal Inference

From Theory to Computation: A Simulation-Based Approach for Difference-in-Differences Validation

Author:
Augusto Gonzalez
Bonorino

December 16, 2023

1 Introduction

Econometric methods for causal inference often grapples with the pervasive issue of endogeneity, particularly in observational data. This is primarily manifested as ignorability, or unconfoundness, assumptions in the potential outcomes framework, where methods often target outcome (e.g., nearest neighbor matching), treatment (e.g., propensity scores), or both (e.g. doubly robust estimation) hoping to adjust for confounding. However, the performance of these methods can vary significantly with finite sample sizes [1], and often impose untestable assumptions about the absence of unobserved confounders. Additionally, the growing interest of leveraging machine learning methods, which require large datasets rare in panel studies, and the challenge of finding valid instruments for 2-stage least squares, highlight the need for more versatile and robust approaches in causal inference. There is a need for more general methods to assess the robustness, consistency, and validity of causal inference models. Popular models like Difference-in-Differences (DiD) are highly context dependent, with their efficacy hinging on the data-generating process, and often relying on extensive institutional knowledge. Traditional testing frameworks share similar requirements.

In an era where computational power and data availability have grown exponentially, there is an emerging opportunity to revisit and innovate these traditional methodologies. The evolution of econometrics is not just a function of theoretical advancement but also a reflection of the technological strides made in computation and data science.

This paper is built on the observation that computational methods are becoming increasingly integral in validating causal inference models. This mirrors trends in complexity economics, where computational methods challenge classical equilibrium models, albeit at the cost of interpretability. The primary aim of this paper is to explore the potential of simulation-based methodologies. These approaches hold the potential to complement and extend traditional econometric analyses, providing a more nuanced and flexible way to examine the robustness of econometric models. While recognizing the limitations of these emerging approaches, their potential remains largely understudied and warrants a deeper investigation.

After this introduction, the paper embarks on an extensive literature review, covering the evolution from traditional econometric methods to sensitivity analysis, and the emerging role of simulation-based approaches in causal inference. A novel taxonomy is proposed to classify methods designed to validate causal assumptions that illustrates the evolution observed in the literature: from traditional, heuristic-based methods to sensitivity analyses grounded in theory, and now, towards more generalizable simulation approaches leveraging cheap computational power and user-friendly programming languages. We then present our methodology for creating synthetic panel data tailored to test the DiD model, detailing a function in pseudocode that addresses selected biases. This leads to a novel taxonomy categorizing methods used in validating causal estimates under untestable assumptions, emphasizing their unique features. Following this, we discuss the methodology, present the results of our experiments, and delve into the implications, limitations, and future directions of simulation-based approaches in econometric research.

2 Literature Review

Causal inference in observational studies is fundamentally challenged by the absence of randomized control, typically deemed the gold standard in establishing causality [2]. While Randomized Controlled Trials (RCTs) offer a high degree of reliability, their implementation is often impractical or ethically questionable in many real-world scenarios [2]; which leaves many causal questions unanswered [3]. Furthermore, results from RCTs face limitations in generalizability to broader contexts [2, 4, 5]. Consequently, econometricians rely on a set of assumptions for identifying causal relationships in observational data [6]. Understanding these assumptions' untestability and their effects is pivotal, as they form the cornerstone of causal inference and its credibility [7, 8]. This is particularly relevant when considering the potential outcomes framework, common in econometric research, which necessitates specific assumptions for causal inference [6]. Moreover, the recent methodological advances in the econometrics of program evaluation highlight the complexity and limitations of these methods in the absence of RCTs [9], leading to the exploration of alternative methodologies in econometrics.

Traditional approaches in econometrics encompass a range of techniques aimed at mitigating biases and validating causal estimates. These methods, including matching algorithms, robustness checks, Manski bounds [10], Lee bounds [11], and multiple regression specifications, are foundational in causal analysis. Notably, matching methods aim to reduce selection bias by approximating random assignment [12]. However, these methods are not without criticism. King and Nielsen's critique of the propensity score's incorrect application in matching [13], and their proposal of alternative approaches that do not rely on balance checking [14], exemplifies concerns regarding model dependency and potential misuse of these methods. Furthermore, traditional approaches often heavily rely on expert intuition, limiting the use of simulations. This reliance categorizes them as face-validity methods, as discussed by Parsikh et al. (2022) [1], underscoring a need for more empirical and simulation-based validation techniques in modern econometric practice. While these traditional methods have been instrumental, the evolving landscape of econometrics has prompted the development of sensitivity analysis techniques, which offer nuanced insights into the robustness of causal inferences.

Sensitivity analysis, pioneered as a method to study the causal relationship between smoking a lung cancer [15], revolves around testing the robustness of conclusions by varying key assumptions. This includes methods like placebo tests and the R^2-parametrization extension of Omitted Variable Bias (OVB) [16], which leverage simulated hypotheses derived from institutional knowledge and theory. Emily Oster's work, particularly her 2019 paper "Unobservable Selection and Coefficient Stability: Theory and Evidence" [17], and its predecessor in 2013 [18], further explores this domain by examining the stability of coefficients in the presence of potential omitted variables. These approaches help quantify "how strong the confounders need to be to invalidate the study's conclusions." VanderWeele and Ding's introduction of the E-Value in 2017 [19] represents another significant advancement in this field, offering a novel metric for assessing the robustness of causal claims against unmeasured confounding. Despite their utility, sensitivity analysis methods require expert knowledge to set plausible scenarios, and their results are often specific to the context of the study, and thus not necessarily generalizable across different models. This evolving field started gaining traction in the early 2000s [3, 20], continually adapting to the complex challenges of causal inference.

Recent advancements in causal inference have been marked by the advent of purely simulation-based approaches, which represent a paradigm shift towards leveraging computational techniques to simulate Data Generating Processes (DGPs). Central to these methods is the use of machine learning, particularly deep generative models [21, 22], to create synthetic datasets that reflect complex real-world scenarios [1]. Susan Athey's work exemplifies this integration, using machine learning for policy evaluation and data generation [23]. Complementary to this, Judea Pearl's Do-calculus provides a theoretical foundation for causal inference in complex systems [7]), while approaches like Bayesian Neural Networks demonstrate the potential of deep learning in uncovering causal mechanisms from intricate datasets [24, 25]. These simulation-based methods, in contrast to traditional synthetic data generators, offer a more nuanced exploration of causal effects under a variety of conditions, showcasing the growing intersection of econometrics with advanced computational techniques.

The increasing integration of computational methods in econometrics does not diminish the value of institutional knowledge, which remains crucial in contextualizing and interpreting causal findings. This paper advocates for a balanced approach where computational techniques are employed to delineate the general limits of causal econometric models, allowing institutional knowledge to be applied more effectively in specific study contexts. The evolution from traditional techniques to contemporary, simulation-based approaches reflects a shift towards more rigorous, transparent, and comprehensive causal inference, enhancing the robustness and credibility of empirical findings in econometrics.

3 Methodology

The utility of simulation-based approaches in testing and validating causal inference models is increasingly recognized in econometrics. These approaches offer a dynamic and customized way to explore the limitations of traditional methods. By simulating various biases and scenarios, researchers can systematically observe their impacts on causal estimates, thus providing insights into the robustness and validity of these models prior to collecting data; arguably the most time consuming aspect of any research project.

3.1 Difference-In-Differences Model

The Difference-in-Differences (DiD) model, a widely used method in econometrics for causal inference where exchangeability cannot be assumed between the treatment and control groups, exemplifies the challenges and complexities inherent in observational studies. DiD is a quasi-experimental design that makes use of longitudinal data from treatment and control groups to obtain an appropriate counterfactual to estimate a causal effect. It is typically used to estimate the effect of a specific intervention or treatment (such as a passage of law, enactment of policy, or large-scale program implementation) by comparing the changes in outcomes over time between a population that is enrolled in a program (the intervention group) and a population that is not (the control group). This paper focuses on the DiD model due to its popularity and because it represents a broader class of models grappling with similar issues in causal analysis under the potential outcomes framework.

By comparing expectations before and after treatment, for both treatment and control groups, the DiD estimator remove biases that could arise from permanent differences between the groups or from external trends affecting the outcomes. Yet, the removal of such biases is not guaranteed. The DiD model, nested within the potential outcomes framework, rests on several critical assumptions, the violation of which can lead to biased and unreliable estimates:

1. Parallel Trends Assumption: The premise that, in the absence of treatment, the outcome trends for treatment and control groups would have been parallel over time.

2. Consistency and No Spillover (part of SUTVA): The treatment received by any unit should not affect the outcomes of other units, and the treatment effects should be consistent across units.

3. Positivity: Ensuring that every unit has a positive probability of receiving each treatment level.

In the selection of biases for simulation within the DiD framework, the guiding criteria were based on two key factors: the likelihood of these biases in empirical research and their consequential impact on DiD estimates. The biases that were chosen for this simulation include unobserved heterogeneity, selection bias, non-parallel trends, and unit heterogeneity. This selection was primarily driven by their documented impact in empirical research, where it has been consistently shown that these biases can significantly affect causal estimates.

3.2 Selected Biases

Furthermore, the sensitivity of DiD estimators to these biases varies; they are generally robust against biases such as selection bias, exhibit moderate sensitivity to time-varying confounders and heterogeneity, and are particularly vulnerable to violations of the parallel trends assumption. By focusing on these biases, the simulation can facilitate a meaningful comparison with empirical findings, thereby providing an opportunity to corroborate the results derived from the simulation approach. This alignment with documented empirical evidence underscores the practical relevance and applicability of the simulation in real-world research scenarios.

Other potential biases, such as SUTVA violations or dynamic treatment effects, were not included due to their lower likelihood in standard DiD applications or the complexity involved in accurately simulating them. The focus on the selected biases offers a balance between empirical relevance and feasibility of simulation, aligning with the primary objective of assessing the most impactful and commonly encountered challenges in DiD models.

3.3 Synthetic Panel Data Generator

This paper introduces a technique designed to generate synthetic panel datasets, allowing for the parametrization of potential sources of bias, particularly those commonly encountered in DiD studies.

The primary objective of this technique is to provide researchers with a flexible tool to explore how different biases might affect their causal inference models. By simulating panel data with customizable bias parameters, the technique offers a practical approach to understanding the impact of specific biases – such as unobserved heterogeneity, selection bias, non-parallel trends, and unit heterogeneity – on DiD estimates. The method's strength lies in its ability to systematically vary these biases, providing insights into their potential impact on real-world analyses. Unlike sensitivity analysis, simulation approaches do not rely on altering or questioning the model's assumptions but rather on creating data or scenarios to "stress test" the model's performance. The objective is to see how the model behaves under various simulated conditions, which helps in understanding the robustness and reliability of the model in practical applications. This implementation serves as an initial step towards a more extensive exploration of biases in causal models, demonstrating the potential of computational simulations in empirical research.

While implemented in R for this study, the underlying methodology is language-agnostic and can be adapted to various computational environments. The pseudocode for this technique is as follows:

```
Function generate_panel_data:
    Input: n_units, n_time, treatment_start_period, treatment_effect,
           selection_bias_intensity, heterogeneity_intensity,
           time_varying_confounder_intensity, trend_difference, noise_sd

    Initialize panel_data with unit and time dimensions
    Define post-treatment period based on treatment_start_period

    Generate observed_covariate as a random normal variable
    Create unobserved_effect based on heterogeneity_intensity
    Simulate time_varying_confounder as a function of time
            and observed_covariate

    Assign treatment based on post period and selection_bias_intensity
    Define control and treatment group trends

    Calculate outcome incorporating treatment_effect, biases, and random noise
    Return panel_data
```

Each parameter in the pseudocode represents a key aspect of the synthetic data, allowing for the simulation of specific biases:

- *n_units*, *n_time*: Define the structure of the panel data. It allows to test the sensitivity towards panel structure (i.e., long, wide, or balanced).
- *treatment_start_period*, *treatment_effect*: Specify the start of treatment and the true treatment effect.
- *selection_bias_intensity*: This parameter sets the intensity of selection bias. It models the degree to which the treatment assignment process is influenced by both observed and unobserved factors, reflecting scenarios where treatment selection is not random.
- *heterogeneity_intensity*: Controls the extent of unobserved heterogeneity among units. This parameter simulates the impact of unobserved variables that may influence both the treatment and the outcome.
- *time_varying_confounder_intensity*: Represents the intensity of time-varying confounders. This parameter is crucial for modeling external

variables that change over time and affect both treatment assignment and outcome.

- *trend_difference*: Adjusts for the presence of non-parallel trends between treatment and control groups. This parameter is key in testing the robustness of the DiD model against violations of the parallel trends assumption.

- *noise_sd*: Specifies the standard deviation of the random noise added to the outcome variable, representing the measurement error or other stochastic variations.

The results section presents the outcomes of applying the proposed algorithm, demonstrating how the simulated biases impact the DiD estimates.

4 Results

This section delineates the outcomes of simulation experiments conducted to evaluate the DiD model's sensitivity to various biases and different panel structures. The model is specified a classical two-way fixed effects implemented with the 'plm' R package. The experiments explore the effects of selection bias, time-varying confounders, parallel trends, heterogeneity bias, and variations in panel length. A deeper analysis then follows, interpreting both the mean estimates and the mean standard errors to provide insights into the model's confidence under various simulated conditions.

Experiment	Mean Estimate	Mean Diff.	Mean SE
Unbiased Balanced Panel	2.983	-0.017	1.212
Unbiased Long Panel	3.001	0.001	1.904
Unbiased Short Panel	3.033	0.033	0.783
Selection Bias	2.998	-0.002	0.548
Time-Varying Confounders	4.756	1.756	0.107
Parallel Trends	28.004	25.004	0.821
Unit-Heterogeneity Bias	3.002	0.002	0.821

Table 1: Summary of Experiment Results

The experiments on panel structures —balanced, long, and short— demonstrate the DiD model's varying performance condintioned on the dataset's structure. Table 1 shows summary results for these experiments, and Figure 1 depicts the results for each of the 50 iterations I ran. All panel structures yield estimates trivially close to the true effect. Interestingly, the mean standard errors vary across these scenarios, with the long panel exhibiting a higher standard error, suggesting less confidence in the estimates. Moreover, all p-values are strongly significant; except the estimates for the unbiased long panel which are significant at the 10% level. Granular results for each iteration, for each experiment, are included in the supplementary R notebook.

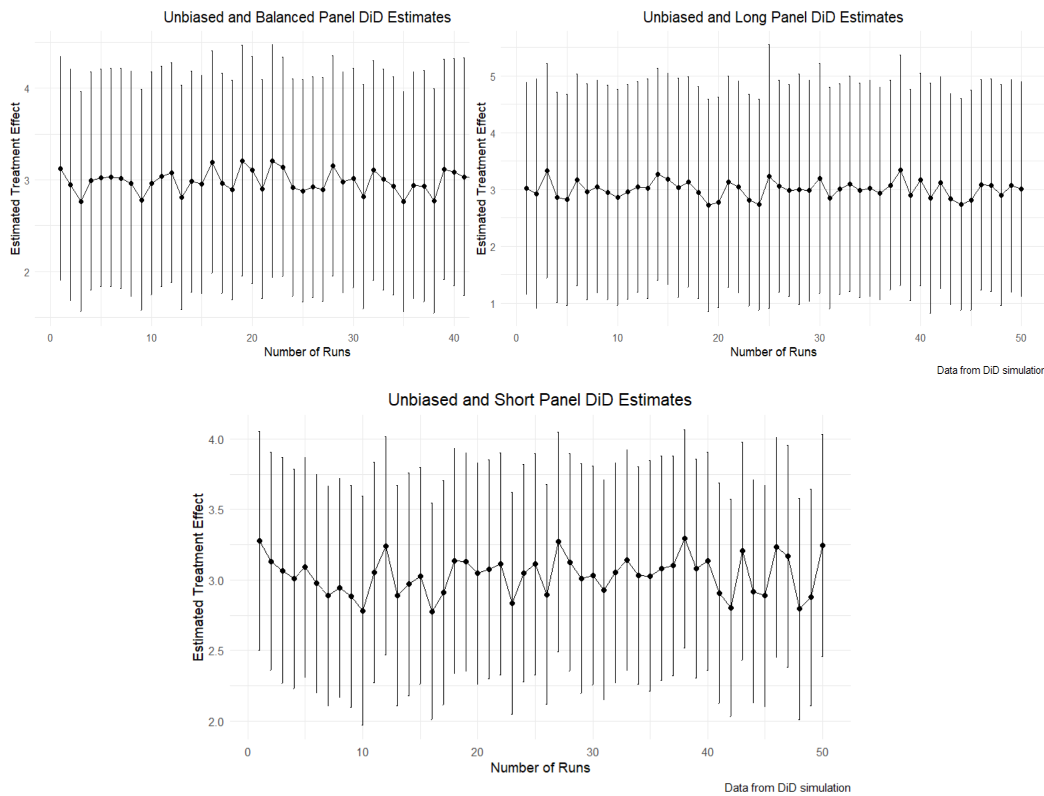

Figure 1: DiD Estimates for Different Panel Structures

The impact of biases on DiD estimates, depicted in Figure 2, is particularly revealing. Selection bias (upper left chart) shows minimal influence on the estimates, evidenced by a negligible mean difference and a relatively low standard error. This reflects the model's robustness to time-invariant selection biases. In contrast, time-varying confounders (lower left chart) and parallel trends (lower right chart) biases cause significant deviations in the estimates. Notably, the parallel trends bias leads to a dramatic increase in the mean estimate but with a surprisingly low standard error. This suggests a deceptive confidence in the estimates, highlighting a potential pitfall in the naive application of the DiD model without thorough pre-analysis testing of assumptions. The heterogeneity bias (upper right chart) exhibits minimal impact, aligning with expectations of the model's ability to handle certain levels of unobserved heterogeneity within units.

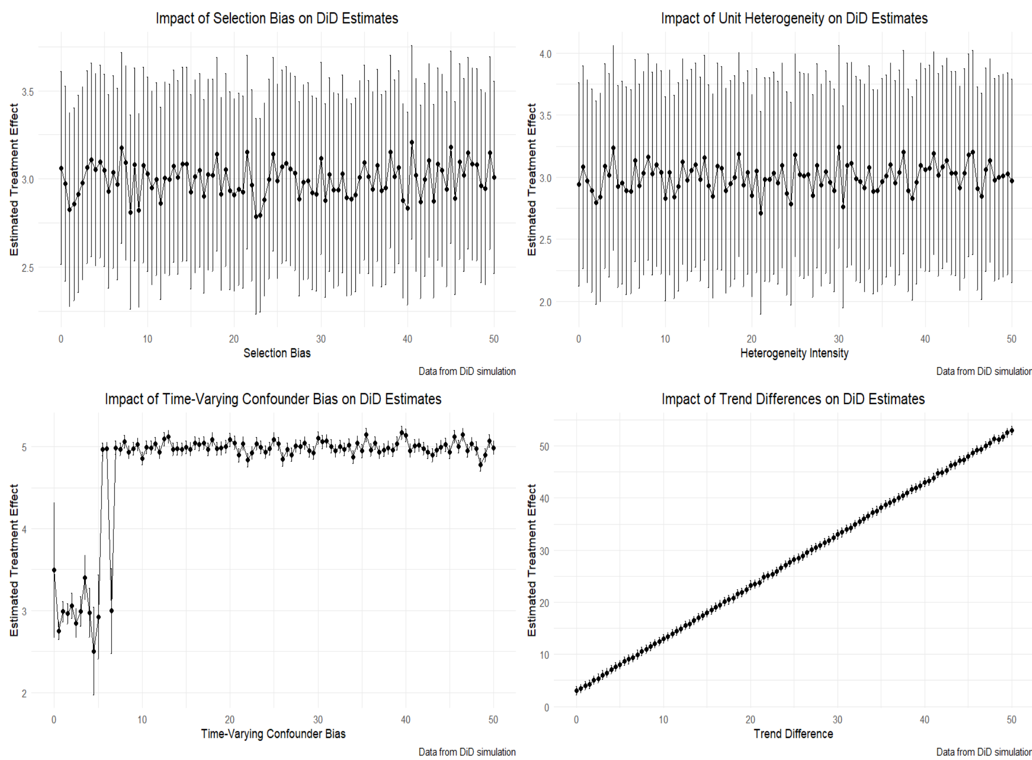

Figure 2: Impact of Various Biases on DiD Estimates

The model's vulnerability to certain biases, especially when coupled with misleading confidence as indicated by low standard errors, underscores the importance of rigorous diagnostic checks and adjusting for standard errors appropriately. These results emphasize the utility of simulation-based approaches in assessing the robustness of econometric models. They provide a framework for understanding the potential biases in real-world scenarios, thereby enhancing the reliability and validity of causal inference analyses in empirical research.

5 Discussion

An extensive review of the literature on causal models supports the taxonomy presented in Table 2:

Methodology	**Identifying Characteristics**	**Advantages**	**Disadvantages**
Traditional Approaches	Employs conventional econometric techniques. Includes within-transformation, TSCS methods, Hausman tests, pre-trend analysis, placebo tests, and event study frameworks. Focuses on indirect testing of model assumptions.	Relatively straightforward and well-understood. Can be implemented with standard statistical software.	Indirect approach; does not conclusively prove assumptions. Can be limited by data availability and quality.
Sensitivity Analysis	Examines the impact of varying key assumptions (like unobserved confounders) on the results. Methods include Manski Bounds, Altonji-Elder-Taber approach, and Oster's approach. Quantifies the robustness of conclusions under different assumption scenarios.	Provides a more nuanced understanding of model dependence. Helps quantify uncertainty due to non-testable assumptions.	Requires complex calculations and assumptions. Results can be difficult to interpret for non-experts.
Simulation Approaches	Involves creating data/scenarios to test model performance. Includes bootstrapping, Monte Carlo simulations, and synthetic control as a simulation tool. Stress tests the model's performance under various conditions.	Directly tests the model's robustness in hypothetical scenarios. Can uncover model limitations not evident in real data.	Requires careful design to ensure realistic simulation scenarios. Computationally intensive and requires technical expertise.

Table 2: Taxonomy of Approaches for Validating Causal Models

The taxonomy in Table 2 provides a structured overview of the methods used in validating causal models, each characterized by distinct features, advantages, and limitations. Traditional approaches rely on conventional econometric techniques and are known for their straightforward application but often fail to conclusively prove assumptions. Sensitivity Analysis methods delve into the effects of varying key assumptions on results, offering a nuanced understanding of model dependence at the cost of complexity and potential interpretation challenges. Our approach, classified under Simulation Approaches, attempts to offer a direct and customizable way to assess the robustness of causal models. Simulation Approaches stand out by creating data or scenarios to directly test model performance, revealing limitations not apparent in real data, though they require additional computational effort. This taxonomy frames our method within the broader spectrum of econometric validation techniques.

Parikh et al. (2022) [1] categorize existing validation approaches into three main categories: face-validity tests, placebo or negative control tests, and synthetic data tests. Our method aligns most closely with the synthetic data tests, but with notable differences in approach and objectives.

Face-validity tests, as discussed by Parikh et al., hinge on expert intuition, a method that, while valuable, is neither sufficient nor necessarily reliable for confirming the accuracy of causal estimates. Placebo tests, widely used in econometrics, are only valid under strong assumptions and can sometimes lead to biased interpretations due to reverse causal-dependency or oversimplified treatment assignment mechanisms. Our approach, while falling under the broader category of synthetic data tests, diverges from traditional implementations in this category by offering user-specified causal treatment effects and heterogeneity, satisfying the first desirable property suggested by Parikh et al. However, the current version does not simulate samples that are stochastically indistinguishable from specific observed data samples, a feature that Parsikh et al. suggest for synthetic data tests. Furthermore, our approach does not address all potential biases that could affect DiD estimates.

6 Conclusion

This paper embarked on an exploratory journey to enhance the tools available for validating causal estimates in econometric models, particularly focusing on the Difference-in-Differences (DiD) approach. By employing a novel simulation-based technique, this research provided insights into how different biases—such as selection bias, time-varying confounders, and violations of parallel trends—impact DiD estimates. The study's distinctive contribution lies in its development and application of a customizable function to generate synthetic panel data, enabling a systematic analysis of these biases. Moreover, the paper introduced a comprehensive taxonomy categorizing existing validation approaches, delineating their characteristics, strengths, and weaknesses.

The findings of this research underscore the critical importance of rigorously testing and understanding the assumptions underpinning causal inference models. The demonstrated sensitivity of DiD estimates to various biases highlights the potential risks in applying these models without due consideration of their limitations. The taxonomy presented serves as a valuable guide for researchers, offering a structured way to select appropriate validation methods depending on the specific challenges and complexities of their empirical studies. This paper's approach to simulation aligns with the growing trend in econometrics to leverage computational power and data science techniques, thereby enhancing the robustness and credibility of econometric analyses.

While the method developed herein marks a significant step forward, it is essential to acknowledge its limitations. The current version of the simulation technique does not mimic the complexity of specific real-world datasets, focusing instead on general scenarios. Future enhancements could aim to integrate more sophisticated, context-specific data-generating processes, thereby increasing the realism and applicability of the simulations. Additionally, expanding the range of biases considered, experiment designs, and exploring the cumulative effects of multiple biases could offer deeper insights into the nuances of causal inference. Exciting avenues for future research include applying this methodology to other econometric models, further bridging the gap between theoretical econometrics and practical data analysis.

In conclusion, this paper reflects a shift in Econometrics towards embracing the power of computation, not just as a tool for analysis, but as an integral component in empirical research. It reaffirms our conviction, as stated in the introduction, that computational methods hold immense potential in extending and refining the methodologies available for testing and validating the assumptions central to causal inference.

In summary, this study contributes to the literature on causal inference validation by offering a novel simulation-based approach. It presents a method that allows researchers to test and understand the limitations of their panel causal models, particularly the DiD model, under various bias conditions. The proposed taxonomy provides a framework for situating different validation methods, including our own, within the broader context of causal inference research. Future enhancements could include expanding the range of simulated biases, allowing for further customization, and exploring ways to tailor the simulations more closely to specific empirical contexts by allowing more complex perturbations.

References

[1] Harsh Parikh, Carlos Varjao, Louise Xu, and Eric Tchetgen Tchetgen. Validating causal inference methods, 2022.

[2] Angus Deaton and Nancy Cartwright. Understanding and misunderstanding randomized controlled trials. Working Paper 22595, National Bureau of Economic Research, September 2016.

[3] Weiwei Liu, Satoko J. Kuramoto, and Elizabeth A. Stuart. An introduction to sensitivity analysis for unobserved confounding in nonexperimental prevention research. *Prevention Science*, 14:570–580, 2013.

[4] Elizabeth Stuart, Catherine Bradshaw, and Philip Leaf. Assessing the generalizability of randomized trial results to target populations. *Prevention science : the official journal of the Society for Prevention Research*, 16, 10 2014.

[5] Antti Malmivaara. Generalizability of findings from randomized controlled trials is limited in the leading general medical journals. *Journal of Clinical Epidemiology*, 107:36–41, 2019.

[6] Donald B. Rubin. Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, 100(469):322–331, 2005.

[7] Judea Pearl. *Causality: Models, Reasoning and Inference*. Cambridge University Press, USA, 2nd edition, 2009.

[8] M.A. Hernan and J.M. Robins. *Causal Inference: What If*. Chapman & Hall/CRC Monographs on Statistics & Applied Probab. CRC Press, 2023.

[9] Guido M Imbens and Jeffrey M Wooldridge. Recent developments in the econometrics of program evaluation. Working Paper 14251, National Bureau of Economic Research, August 2008.

[10] Charles F. Manski. Nonparametric bounds on treatment effects. *The American Economic Review*, 80:319–323, 1990.

[11] David S Lee. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. Working Paper 11721, National Bureau of Economic Research, October 2005.

[12] Kosuke Imai, In Song Kim, and Erik H. Wang. Matching methods for causal inference with time-series cross-sectional data. *American Journal of Political Science*, 67(3):587–605, 2023.

[13] Gary King and Richard Nielsen. Why propensity scores should not be used for matching. *Political Analysis*, 27(4):435–454, 2019.

[14] Stefano M. Iacus, Gary King, and Giuseppe Porro. Causal inference without balance checking: Coarsened exact matching. *Political Analysis*, 20(1):1–24, 2012.

[15] Jerome Cornfield, William Haenszel, E. Cuyler Hammond, Abraham M. Lilienfeld, Michael B. Shimkin, and Ernst L. Wynder. Smoking and lung cancer: recent evidence and a discussion of some questions*. *International Journal of Epidemiology*, 38(5):1175–1191, 09 2009.

[16] Carlos Cinelli and Chad Hazlett. Making Sense of Sensitivity: Extending Omitted Variable Bias. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 82(1):39–67, 12 2019.

[17] Emily Oster. Unobservable selection and coefficient stability: Theory and evidence. *Journal of Business & Economic Statistics*, 37(2):187–204, 2019.

[18] Emily Oster. Unobservable selection and coefficient stability: Theory and validation. Working Paper 19054, National Bureau of Economic Research, May 2013.

[19] Tyler J. VanderWeele and Peng Ding. Sensitivity analysis in observational research: Introducing the e-value. *Annals of Internal Medicine*, 167:268–274, 2017.

[20] Kenneth A. Frank, Qinyun Lin, Ran Xu, Spiro Maroulis, and Anna Mueller. Quantifying the robustness of causal inferences: Sensitivity analysis for pragmatic social science. *Social Science Research*, 110:102815, 2023.

[21] Debo Cheng, Ziqi Xu, Jiuyong Li, Lin Liu, Jixue Liu, and Thuc Duy Le. Causal inference with conditional instruments using deep generative models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(6):7122–7130, Jun. 2023.

[22] Tomas Geffner, Javier Antoran, Adam Foster, Wenbo Gong, Chao Ma, Emre Kiciman, Amit Sharma, Angus Lamb, Martin Kukla, Nick Pawlowski, Miltiadis Allamanis, and Cheng Zhang. Deep end-to-end causal inference, 2022.

[23] Susan Athey. Beyond prediction: Using big data for policy problems. *Science*, 355(6324):483–485, 2017.

[24] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. GAIN: Missing data imputation using generative adversarial nets. In Jennifer Dy and Andreas Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 5689–5698. PMLR, 10–15 Jul 2018.

[25] Jinsung Yoon, James Jordon, and Mihaela van der Schaar. Ganite: Estimation of individualized treatment effects using generative adversarial nets. In *International Conference on Learning Representations*, 2018.