CARMA 2026

8th Int. Conf. on Advanced Research Methods and Analytics
1-3 July 2026, Valencia, Spain
DOI: https://doi.org/10.4995/CARMA2026.2026.21912

Is TWFE Enough? When to Switch to Double Machine Learning in Differences in Differences

Martin Gabriel Cargnel[1], **Maria Noelia Romero**[2]

[1]Escuela de Negocios y Administración Pública, Facultad de Ciencias Económicas, Universidad de Buenos Aires, Argentina, [2]Departmento de Economía y Centro de Estudios para el Desarrollo Humano (CEDH), Universidad de San Andrés, Argentina.

How to cite: Cargnel, M.G.; Romero, M.N. (2026). Is TWFE Enough? When to Switch to Double Machine Learning in Differences in Differences. In: 8th International Conference on Advanced Research Methods and Analytics (CARMA 2026). Valencia, 1-3 July 2026. https://doi.org/10.4995/CARMA2026.2026.21912

Abstract

While Machine Learning (ML) excels at prediction, causal inference has traditionally relied on simpler models. Double Machine Learning (DML) bridges this gap, leveraging flexible ML algorithms for robust causal estimation. The primary aim of this work is to establish when DML is preferred over classical TWFE. We compare both approaches using two empirical cases: a standard setup where both methods yield similar estimates, and a complex scenario where TWFE produces counter-intuitive results while DML maintains accuracy. Our findings demonstrate that DML successfully isolates causal effects where traditional linear estimators fail. We conclude that while TWFE suffices for standard setups, DML provides a necessary, robust alternative for unbiased causal inference in complex DiD environments.

Keywords: *Double Machine Learning; Differences in Differences; Causal Inference; TWFE.*

1. Introduction

When working in applied statistics, there is a fundamental distinction between predicting outcomes and explaining relationships (Shmueli, 2010). Highly flexible machine learning (ML) algorithms excel at maximizing predictive performance, but their black-box nature often precludes the estimation of causal relationships (Mullainathan & Spiess, 2017). Consequently, economists and social scientists have historically favored simpler, interpretable approaches like linear regressions to estimate causality.

While recent advancements in interpretable ML address model transparency (Molnar, 2025), they do not inherently solve the challenge of uncovering the underlying data-generating process with formal statistical guarantees. To bridge the gap between predictive ML and causal

Article in Press

This work is licensed under a Creative Commons License CC BY-NC-SA 4.0
Editorial Universitat Politècnica de València

inference, the Double Machine Learning (DML) framework (Chernozhukov et al., 2018) emerged. DML utilizes machine learning not to explain predictions, but to flexibly control for high-dimensional confounding variables while maintaining valid statistical inference on targeted causal parameters.

This paper examines the application of DML within the Difference-in-Differences (DiD) framework. We evaluate DML against traditional TWFE estimators in both a canonical simultaneous adoption setting and a complex staggered adoption scenario to demonstrate when DML is preferred for unbiased causal estimation. Our contribution is practitioner-oriented: we provide a case-based diagnostic for when the additional complexity of DML is needed over classical TWFE, and when it is not. As the set of available estimators continues to grow, such guidance helps applied researchers match the method to the empirical setting.

2. Methodology

2.1. The Traditional DiD and TWFE Limitations

DiD estimates causal effects by comparing changes in outcomes over time between a treated group and an untreated control group.

The traditional approach often relies on a TWFE regression:

$$Y_{it} = \alpha_i + \lambda_t + \delta D_{it} + \epsilon_{it} \quad (1)$$

where Y_{it} is the outcome, α_i and λ_t are unit and time fixed effects, D_{it} is the treatment indicator, and δ represents the average treatment effect. While TWFE is robust in simultaneous adoption settings, it often produces biased estimates under staggered adoption designs (Goodman-Bacon, 2021). Recent advances in the DiD literature have proposed estimators that address some of these limitations, most prominently the cohort-time ATT framework of Callaway & Sant'Anna (2021) and the interaction-weighted estimator of Sun & Abraham (2021). However, they are typically applied with parametric specifications for conditioning on covariates X_{it} , which risks bias when the true data-generating process involves complex, nonlinear interactions.

2.2. The Double Machine Learning Framework

DML addresses confounding without imposing strict functional forms. The objective is to estimate the Average Treatment Effect (ATE), represented by θ, within a partially linear model:

$$Y_i = \theta D_i + g(X_i) + \epsilon_i \quad (2)$$

Article in Press

$$\boldsymbol{D_i = m(X_i) + u_i} \quad \textbf{(3)}$$

Here, $g(X_i)$ and $m(X_i)$ are unknown, complex nuisance functions that represent how covariates X affect the outcome and treatment, respectively.

The model orthogonalizes the variables, reducing the complex system to a simple linear regression of residualized outcomes on residualized treatments, leveraging flexible ML algorithms (e.g., gradient boosting, random forests) to estimate the conditional expectations required to compute these residuals.

To ensure valid inference, DML relies on two critical techniques: (1) Sample splitting ensures models are trained on different data folds than those used to generate residuals, eliminating overfitting bias and (2) Neyman Orthogonality prevents ML regularization bias from contaminating the causal estimate θ.

2.3. DML for DiD Settings

For a canonical DiD, the framework utilizes a doubly robust Neyman-orthogonal score function, requiring cross-fitted ML estimates of both the propensity score and the expected counterfactual outcome change (Chang, 2020). For staggered adoption, the framework estimates cohort-specific treatment effects, $ATT(g,t)$, comparing treated units of cohort g against a control group of never-treated or not-yet-treated units. These specific effects are then aggregated using relative size weights to produce an overall dynamic event study or average treatment effect (Callaway & Sant'Anna, 2021).

2.4. DML Implementation details

The DML-DiD estimators are implemented in Python using DoubleML (Bach et al., 2021). Cross-fitting follows the standard K-fold approach with multiple repetitions to reduce sample-split variability: K=5 with 10 repetitions for the canonical application, and K=2 with 10 repetitions for the staggered application to accommodate the smaller state-year panel and the per-cohort estimation cost. For each application, nuisance functions are estimated separately for the propensity score and for outcome regressions on treated and control subsamples, with hyperparameters tuned via Optuna optimization (50 trials) using cross-validation on the training folds. Base learners are chosen to balance flexibility against the dimensionality of the conditioning set: gradient boosting trees (Ke et al., 2017) for the canonical application, with two low-dimensional covariates, and random forests (Breiman, 2001) for the staggered application, with a higher-dimensional set of demographic, economic, and policy controls. In the staggered design, the 2005 and 2009 adoption cohorts (one state each) are excluded from

Article in Press

them estimation, since the doubly robust score requires sufficient within-cohort observations for stable nuisance estimation[1].

3. Empirical Applications

We compare TWFE and DML-DiD estimators across two empirical settings: a canonical single-period treatment application and a complex staggered adoption application.

3.1. Single-Period Treatment: Fracking Regulation

We evaluate the effect of the 2005 fracking boom on environmental regulatory activities using zip-code-year level data, following the approach and data from González (2025). Because treatment occurred simultaneously across all affected units, this represents a classic DiD design.

Both TWFE and DML-DiD estimators yield positive and statistically significant results for all three outcomes, as shown in Figure 1. For Actions, TWFE estimates a coefficient of 0.229 while DML yields 0.193; for Facilities, TWFE estimates 0.159 and DML 0.127; and for Formal Activities, TWFE and DML produce 0.209 and 0.202, respectively. While DML confidence intervals are noticeably wider than those from TWFE, reflecting the inferential cost of relaxing parametric functional-form assumptions, the qualitative conclusions are identical across methods. These results indicate that in simple setups with parsimonious covariates, traditional linear specifications accurately capture the causal effect at lower inferential cost.

3.2. Staggered Adoption: Castle Doctrine Laws

Next, we analyze the impact of Castle Doctrine laws (which expand the right to use deadly force in self-defense) on state-level homicide rates from 2000 to 2010. Because states adopted these laws in different years, this presents a complex staggered adoption landscape.

While Cheng & Hoekstra (2013) found a positive significant effect in this context, our TWFE specification produces a point estimate of 0.087 with wide confidence intervals (-0.032 to 0.206), suggesting a non-significant effect that fails to replicate the established literature. In contrast, the DML-DiD estimator successfully recovers a positive, significant effect of 0.131 (CI: 0.040 to 0.216), aligning with prior empirical evidence. These results are summarized in Figure 2.

Article in Press

[1] Complete implementation details, hyperparameters search ranges and replication code are available at https://github.com/mcargnel/mea

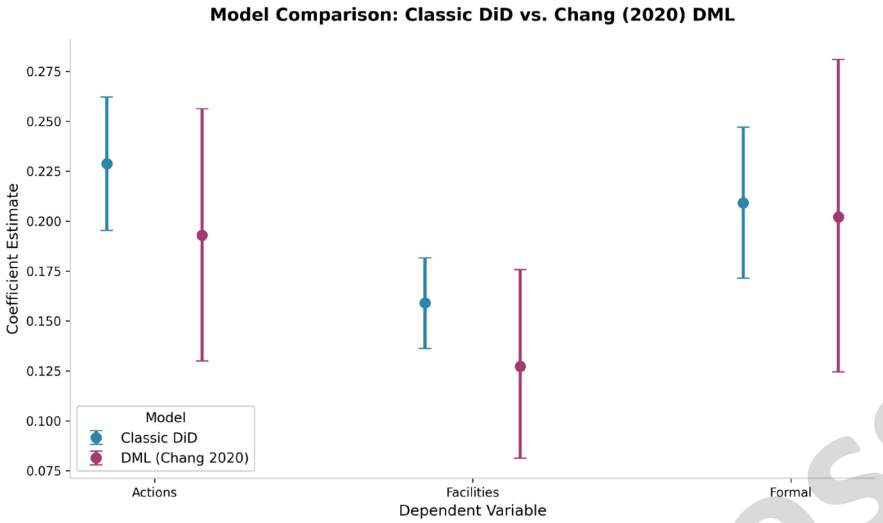


Figure 1. Treatment effect estimates for the impact of fracking on environmental regulation outcomes using TWFE (blue) and DML (violet) estimators. Point estimates and 95 percent confidence intervals are shown for three outcomes: Actions (total regulatory activities), Facilities (facilities receiving actions), and Formal (formal enforcement activities). All outcomes use log transformations. Both estimators yield positive and statistically significant point estimates across all three outcomes. Source: Author's elaboration.

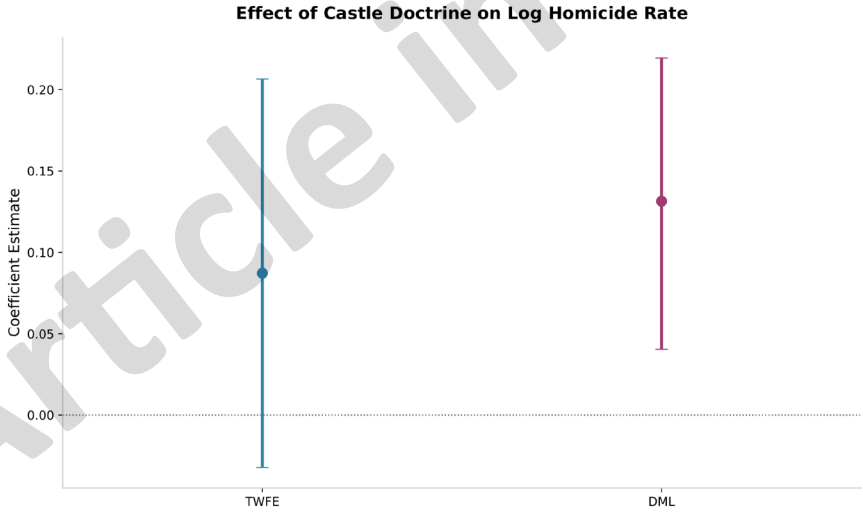


Figure 2. Comparison of the estimated effects of Castle Doctrine laws on homicide rates using the traditional TWFE estimator and the Double Machine Learning DiD estimator adapted for staggered adoption settings. The confidence intervals are represented for both methods, and the horizontal line at zero indicates no effect. Source: Author's elaboration.

Article in Press

Event study analysis (Table 1) assesses the parallel trends assumption and the dynamic evolution of treatment effects for Castle Doctrine laws on log homicide rates. Using our DML-DiD framework, we observe statistically significant pre-treatment coefficients at $t = -6$ and $t = -1$. However, rather than exhibiting a systematic, monotonic divergence that would threaten design validity, these estimates fluctuate around zero. This suggests the significance is likely driven by idiosyncratic state-year shocks rather than a fundamental violation of parallel trends. Furthermore, the significance at $t = -1$ may capture anticipation effects occurring shortly before enactment.

Post-treatment, a consistent pattern emerges: significant positive effects appear at $t = 0$ and persist in years 1, 2, and 4. Unlike the pre-period noise, these post-treatment coefficients are consistently positive, firmly reinforcing our main finding that Castle Doctrine laws increase homicide rates.

Table 1. Event Study Estimates of Castle Doctrine Laws on Log Homicide Rates. Notes: This table reports dynamic treatment effects by years relative to law adoption, estimated via Double Machine Learning Difference-in-Differences (DML-DiD). The columns display point estimates, standard errors, and 95% confidence intervals. Significance levels for pre-trends and post-treatment effects are denoted as: * p < 0.10, ** p < 0.05, *** p < 0.01. Source: Author's calculations.

Years	Coeff	Std. Err.	2.5%	97.5%
-7 years	-0.060	0.087	-0.261	0.111
-6 years	0.230**	0.093	0.050	0.412
-5 years	-0.014	0.063	-0.137	0.111
-4 years	-0.030	0.064	-0.148	0.096
-3 years	0.001	0.063	-0.102	0.124
-2 years	0.043	0.047	-0.044	0.135
-1 years	-0.111**	0.052	-0.220	-0.008
0 years	0.125**	0.050	0.028	0.222
1 years	0.096*	0.062	-0.016	0.218
2 years	0.128**	0.063	0.008	0.251
3 years	0.096	0.078	-0.050	0.249
4 years	0.183**	0.077	0.027	0.335

4. Conclusions

In this study, we demonstrated that integrating flexible machine learning into the Difference-in-Differences (DiD) framework provides a robust toolkit for causal inference in complex empirical settings. In canonical designs with simultaneous treatment and limited covariates, traditional TWFE and Double Machine Learning (DML) yield comparable results. However, in staggered adoption scenarios, DML successfully accounts for dynamic treatment effects and high-dimensional confounders where TWFE produces biased, non-significant estimates.

Article in Press

Ultimately, DML acts as an essential complement to economic intuition, enhancing the credibility of causal estimates within properly defined identification strategies.

References

Bach, P., Chernozhukov, V., Kurz, M. S., & Spindler, M. (2021). DoubleML -- An Object-Oriented Implementation of Double Machine Learning in Python (arXiv:2104.03220). arXiv. https://doi.org/10.48550/arXiv.2104.03220

Breiman, L. (2001). Random Forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Callaway, B., & Santa'Anna, P. H. C. (2021). Difference-in-Differences with multiple time periods. Journal of Econometrics, 225(2), 200–230. https://doi.org/10.1016/j.jeconom.2020.12.001

Chang, N.-C. (2020). Double/debiased machine learning for difference-in-differences models. The Econometrics Journal, 23(2), 177–191. https://doi.org/10.1093/ectj/utaa001

Cheng, C., & Hoekstra, M. (2013). Does Strengthening Self-Defense Law Deter Crime or Escalate Violence? Evidence from Expansions to Castle Doctrine. The Journal of Human Resources, 48(3), 821–853.

Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018). Double/debiased machine learning for treatment and structural parameters. The Econometrics Journal, 21(1), C1–C68. https://doi.org/10.1111/ectj.12097

González, J. P. (2025). Environmental regulation, regulatory spillovers and rent-seeking. Public Choice, 202(1), 217–250. https://doi.org/10.1007/s11127-024-01189-7

Goodman-Bacon, A. (2021). Difference-in-differences with variation in treatment timing. Journal of Econometrics, 225(2), 254–277. https://doi.org/10.1016/j.jeconom.2021.03.014

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: a highly efficient gradient boosting decision tree. Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS'17, 3149–3157.

Molnar, C. (2025). Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (3rd ed.). https://christophm.github.io/interpretable-ml-book

Mullainathan, S., & Spiess, J. (2017). Machine Learning: An Applied Econometric Approach. Journal of Economic Perspectives, 31(2), 87–106. https://doi.org/10.1257/jep.31.2.87

Shmueli, G. (2010). To Explain or to Predict? Statistical Science, 25(3), 289–310. https://doi.org/10.1214/10-STS330

Article in Press