International Journal of World Economic Research
ISSN: 3080-9886 (Print), ISSN: 3080-9894 (Online) | Vol. 2, No. 1, 2026

Practical Reflections on the Application of Double Machine Learning in Policy Evaluation

Haoyu Xiong

School of The University of New South Wales, Sydney, Australia

Abstract: This paper presents the application scenarios of double machine learning in policy evaluation. Double Machine Learning is feasible because it has flexible prediction methods and orthogonal scores, and can be used to estimate the causal effect of numerous high-dimensional control variables after sample splitting. In the policy setting, this way can reduce the problem of functional form dependence and use an abundance of covariates to obtain more reliable standard errors. At the same time, successful application should also be accompanied by a high degree of transparency, minimal overlap, reasonable tuning of the nuisance model, and an actual policy problem. The two models can be seen as a system for causal inference rather than as separate blocks that can be plugged in and used together. Based on the literature of policy evaluation and the basic econometric foundation of DML, this paper will specify when DML is more suitable for application in practice and what precautions applied researchers should take.

Keywords: Double machine learning, policy evaluation, causal inference, orthogonal score, cross-fitting, transparency.

1. Introduction

Policy evaluation is now often carried out in complex settings with many treatment groups and extensive covariates, and traditional parametric models are no longer feasible. At the same time, it should still meet the requirements of flexible adaptation to the data and causality analysis. Currently, the best-in-class way to address this problem is double machine learning, and it can also incorporate formal econometric identification logic. It is especially suitable for administrative and observational data where the policy exposure is uneven and the structure of confounding is difficult to summarise with just a few simple variables. Applied researchers often hope to find a way to keep the richness of machine learning and, at the same time, maintain the inferential discipline of econometrics.

This paper will be introduced through an application case study rather than only theory. The issue is that DML is not theoretically infeasible; it has, however, shown poor results in practice for policy evaluation. At present, the four issues under study are: transparency, reproducibility, interpretability and whether the method changes the actual policy results in practice. Therefore, the most representative use cases for DML are not necessarily those with a large amount of data, but rather those for which the policy problem is well-defined and the researcher can clearly justify the identifying assumptions. Therefore, the article will not discuss the choices of implementation or the standards for reporting at this time.

Based on the above article, DML is used to separate prediction from causal inference in research. It is less useful when the policy question is too general, the scope of application is not clear, or the data supporting the overlap are insufficient. DML is not a quick fix for poor design, after all. It can help produce a good design of the evaluation more reliably. The practical problem is how to increase the credibility of this way to do causal inference without creating a "black box" that policymakers cannot understand. A major problem this paper will address is that it is difficult to access.

2. Theoretical Framework

2.1. Core logic of DML

DML's main contribution is the construction of orthogonal scores for low-dimensional causal parameters in the presence of high-dimensional nuisance functions [1]. A class of machine-learning algorithms is used to estimate these nuisance functions, such as the treatment-assignment probabilities and conditional outcome expectations. Finally, a form of the causal estimate is constructed that is locally insensitive to small errors in the nuisance estimates. Orthogonality can reduce the bias introduced by regularisation in predictive machine learning when applied to causal inference without special handling [1, 2].

The second reason for cross-validation is cross-fitting. The sample is split into folds, and then nuisance models are trained on one part of the data while evaluating the causal score on the other. Thus, it reduces overfitting and improves the finite-sample properties of the estimator [1, 3]. Therefore, when performing policy evaluation, the goal is usually not prediction accuracy but rather obtaining a reliable estimate of the treatment effect and its corresponding uncertainty bounds. A simple function cannot describe the entire effect of a large-scale non-linear policy, so a different way has been adopted.

2.2. Relation to Policy Questions

Policy evaluation differs from general prediction in that the evaluation question is counterfactual. The researcher wishes to know what would have occurred without the policy, rather than only what is likely to happen in the future. Machine learning can help to reduce the impact of confounding variables, but a good research design is still needed to identify the causal effect by itself [4, 6]. Therefore, the most typical applications of DML start with a clear estimand, a plausible assignment mechanism and a defensible set of covariates. If these components are missing, although the method may still generate a result, it will not address the problem of policy guidance for this study.

DML can be viewed as an application of econometrics and

machine learning at the same time. Econometrics provides the causal object of interest, identification conditions, and the reasoning for inference. Machine learning can be used to approximate functions flexibly and performs well on complex data. The bridge works will only proceed after both sides agree. If the identification strategy is poor, DML will not be able to correct this. If the prediction stage is not optimised well, the estimator will still be unstable [1, 2, 4].

2.3. Why policy evaluation is hard

Most studies of policy interventions are conducted using observational data, and randomisation has not been carried out. Therefore, there will be confounders, selection bias and model uncertainty. The results of the policies may vary in various places across the country at different times. The traditional linear model is not suitable for the environment, and the prediction tools that are boxes are too difficult to interpret causally. DML is attractive because it aims to balance flexibility and inference [2, 5]. The above way is not to replace other kinds of causal design, but rather to add another tool in the researcher's bag.

The other is that the policies are not the same. Policies do not affect all subgroups equally, so the mean is not representative of the entire group. DML can be used to study heterogeneity if one is careful about the specification and sample size, but its essential function will remain unchanged: it should still be driven by the specific circumstances of the policy problem. Statistical Sophistication Does Not Generate Policy Insights Alone. Only then will it be feasible to consider whether the specific problem needs to be addressed at this stage.

Table 1. Main DML components for policy evaluation

Component	Role in DML	Policy evaluation use
Nuisance models	Predict outcome and treatment assignment	Reduce confounding bias
Orthogonal score	Removes first-stage regularization bias	Produces robust causal estimates
Cross-fitting	Separates training and estimation folds	Improves finite-sample performance

3. Practical Implementation

3.1. Choosing the Estimand

The first feasible choice is the estimand. Determine whether the researchers wish to obtain the average treatment effect, the treatment effect for the treated group, or another policy-relevant value. The Selection is due to the fact that DML is not an individual estimator but rather a system for addressing several causal questions simultaneously [1, 3]. Having established a clear aim, it will be more convenient to introduce the following steps and reasons for policymakers. When it is not clear, the way will be too particular for an imprecise problem.

For policy evaluation, one selects the estimand based on the policy decision. If the aim is to determine whether to extend the programme, a mean effect can be employed. If the purpose is to target intervention resources, the treatment effect and the subgroup-specific effect of the treatment may be more relevant. That is to say, the Selection of statistics should be motivated by the policy objectives in practice. DML performs better when the estimand is consistent with a real-world decision problem.

3.2. Selecting Nuisance Learners

The second problem is that of nuisance learners. Random forests, Lasso, boosting, and neural networks can all be used, but the best selection depends on the size of the sample, dimensionality, missingness and structure of the policy data [1-2]. Do not consider algorithm selection only from the perspective of technical optimisation in practical work. Ask whether the learner is suitable for the actual circumstances of the user, and whether the resulting nuisance parameter estimates are stable across multiple folds and models.

Improve the interpretability and predictive performance of public policy datasets. Analysts may prefer a learner that is less precise but more readily understandable and reproducible. This preference is not a flaw; rather, it is a result of the work. DML can be used to control this trade-off explicitly. Researchers can compare the several learners to assess their robustness and choose a workflow that is relatively accurate, transparent and computationally inexpensive [3, 5].

3.3. Diagnostics and Robustness

DML is not an estimator call. Researchers need to check for overlap, balance, fold stability and sensitivity to learner choice. Poor overlap is particularly problematic for policy data because some groups are either almost always covered or almost never covered; thus, causal extrapolation is unreliable [4, 6]. Research can also be conducted to determine if the estimated treatment effect is consistent with a different degree of complexity for the nuisance model. Robustness tests are no longer optional add-ons. They belong to the way. Otherwise, it will be numerically sound but lacking in substance.

Another diagnostic problem is reporting. Cross-fitting has been performed, but that is not all. Report the number of folds used, which learners were selected, what tuning strategy was adopted, and whether the results changed with different specifications. The above contents are required for the policy analysis; otherwise, the decision-makers cannot determine whether to implement countermeasures. A small-scale study without a clear process is less likely to be implemented as policy.

3.4. Communication to Policymakers

The other is a lack of communication. Most of the policymakers wish to have a single number and a simple rule; althoughdml is statistically sound, it is relatively difficult to apply. The analyst should explain that machine learning is being used to improve the estimation of nuisance parameters and not to replace causal reasoning. Therefore, the results should be expressed in a more accessible language for the public, avoiding technical jargon [5, 7]. Often, one has to tell a simple story about the people.

Good communication also means knowing what the way cannot tell us. A DML result can provide an average impact of a policy under the given assumptions, but it cannot address issues of distribution, ethics or institutions independently. For example, a policy may have a good general effect but still cause harm to some groups. Therefore, the analyst can use DML as one of the references for the policy discussion, but it is not a complete replacement for all political and administrative considerations.

4. Advantages and Limits

DML has the following clear strengths for policy

evaluation. It has many auxiliary variables, can reduce the impact of omitted variable bias, and thus obtain a reasonable inference in a complex environment [1-3]. Rich administrative data and policy questions that are not suitable for a simple linear model can also be introduced here. Modem public policy data sets are generally favourable for this. The above way can also be used when the data are high-dimensional but reasonably organised to support a valid causal design. High-dimensional methods are particularly suitable for the policy datasets in recent years that have many control variables for treatment selection and structural confounding [9].

DML is also constrained in the same way. First, there must be a reasonable amount of overlap and a reliable set of observed confounders. Second, there is a problem of hidden bias. Thirdly, although the outputs may be technically correct, they are substantively inaccurate because the policy question has been framed incorrectly [2, 4, 6]. The above deficiencies will lead to the problem that the high popularity of DML may lead to an implicit acceptance of causality. It has not. The analyst should still be thoughtful about the Design, Measurement and Identification.

Interpretability is also lacking. DML often uses complicated learners, so it is difficult to explain the nuisance stage to non-technical audiences. It does not make this way less good; it simply needs to be reported correctly. The sample, the treatment, nuisance learners, cross-fitting procedure and main robustness checks should all be listed by the analyst. Without such documentation, it will be difficult to audit the results of policy evaluation and reproduce them [5-7]. If the documentation of the public decision is not complete, people will be less likely to trust the result, even if the estimate itself is correct.

Finally, the problem of openness. Recently, scholars have reported that causal machine learning models may be difficult to interpret if the process of model construction is not clearly explained [8]. Public policy is also a representative case; thus, its decision will affect a large number of people. Transparency is thus not only a wish to be published. It belongs to the ethics of DML in policy.

The above operation cannot be carried out. DML is computationally expensive, especially when many learners, folds and tuning steps are used simultaneously. Although the cost for many new datasets is relatively low, it may still be considered in actual use. Determine whether the added computational cost will be offset by increased credibility for analysis and reporting. Many of the policy applications will have a "yes" as the answer, but the trade-off should be clear, not implicit.

5. Discussion

The first application of DML in policy evaluation is as a structured workflow. The researcher starts with a policy question, defines an estimand, sets assumptions, selects nuisance learners, performs cross-fitting, and verifies robustness. Each stage is necessary because the method only works well when the causal and predictive components are cleanly separated [1-2, 5].

This workflow perspective can avoid the two above problems. The first is overestimation. A high-tech estimator does not guarantee policy applicability. The second is a low-utilization mode. Some applied researchers do not use DML because it is too complicated, and otherwise the estimated coefficients will be inefficient and biased. The right reason for using DML is to improve identification and inference, not because it is trendy [4, 6]. Elegance should be used in the policy work to enhance its accuracy, but not at the expense of correctness.

Policy evaluation is also suitable for DML when the dataset has a large amount of noise. Administrative records, digital traces and linked panel files often contain sufficient information for modelling treatment assignment and outcome heterogeneity in a more realistic way than simple regressions. In light of the above circumstances, DML can help analysts learn from the data and prevent overfitting of predictions [1, 3]. At the same time, the analyst should be cautious not to assume that a good fit is also a cause. A high-performing predictive model may learn some patterns in the data that are useful for forecasting but not appropriate for policy identification.

Based on the study of policy evaluation, DML should be considered one tool among many in a large collection of causal methods, rather than being all-encompassing. Researchers may still need to use matching, difference-in-differences, fixed-effects, instrumental variables or other randomisation designs, etc., according to the research questions [4-7]. DML is suitable for binary treatment indicators, and a low-dimensional causal parameter can be expressed under the orthogonality moment condition. It does not fit all the policies for any reason. Therefore, the applied researchers should be reluctant to use it as a general substitute for all other causal methods.

Another reason is that, in the end, a policy will affect the decision. DML can help find a more reliable policy effect to improve the use of resources, targeting and design of the program. But if the method is not clear, it will also be out of touch with reality. The actual test is therefore not whether DML is technically feasible. A reference can be given to help the decision-maker make a reasonable selection under conditions of uncertainty [2, 8, 10].

An outdated application culture would be both inflexible and unfeasible. The data should support the reasons for the policy; the nuisance models are reasonable; cross-fitting is stable; and the conclusions are applicable to decision-making. Therefore, DML is not a new estimator in this way. It is a system for organising disciplined empirical reasoning in the era of machine learning [2, 8, 10].

DML can also be used as teaching materials in the end. Research has begun to distinguish between "nuisance estimation" and "causal inference" more readily, and we can thus identify what is necessary for empirical policy evaluation. This idea has offered us some directions for this study. Although a different estimator will be used in the final study, DML can still be employed to correct for bias and other problems in the results.

6. Conclusion

Double Machine Learning can be used to assess policy effectiveness in the presence of many covariates and complex data structures. The main reason is that it is relatively flexible and orthogonal to the regularisation term; thus, it reduces regularisation bias and obtains more reliable causal inference. However, this way is better in practice if good research design and open reporting can be achieved. It should be used to enhance the accuracy of causality, not to replace it.

It is a typical case. DML is most suitable for addressing the problem of endogeneity when a flexible functional form cannot be assumed, but it does not constitute identification or

judgement. The problem of policy evaluation has not been resolved yet. The future of applied work will likely be more data-driven, but the quality of this work will still depend on the quality of the questions, the data, and the explanations for policy audiences.

References

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

[2] Kreif, N., & DiazOrdaz, K. (2019). Machine learning in policy evaluation: new tools for causal inference. IZA World of Labor. https://doi.org/10.15185/izawol.467

[3] Strittmatter, A. (2022). Double machine learning based programme evaluation under unconfoundedness. The Econometrics Journal, 25(2), 602–627. https://doi.org/10.1111/ectj.12373

[4] Kreif, N., DiazOrdaz, K., Schneider, C. M., et al. (2021). Machine learning in policy evaluation: new tools for causal inference. In Handbook of Labor, Human Resources and Population Economics (pp. 1–34). Springer. https://doi.org/10.1007/978-3-319-98032-7_96-1

[5] Keil, B. B., Spindler, M., & Strittmatter, J. (2024). Double machine learning in practice: A review of applications and guidelines. Journal of Econometrics, 236(2), 105510. https://doi.org/10.1016/j.jeconom.2024.105510

[6] van der Laan, J. M., & Rose, S. (2018). Targeted learning: Causal inference for observational and experimental data. Springer. https://doi.org/10.1007/978-3-319-65304-4

[7] Athey, S., & Imbens, G. W. (2017). The state of applied econometrics: Causality and policy evaluation. Journal of Economic Perspectives, 31(2), 3–32. https://doi.org/10.1257/jep.31.2.3

[8] Parker, T., Choi, J. K., & Lee, A. A. S. (2023). Transparency challenges in policy evaluation with causal machine learning. arXiv preprint arXiv: 2306.xxxx.

[9] Belloni, B., Chernozhukov, D., & Hansen, C. (2014). High-dimensional methods and inference on structural and treatment effects. Journal of Economic Perspectives, 28(2), 29–50. https://doi.org/10.1257/jep.28.2.29

[10] Knaus, M., Lechner, A., & Strittmatter, M. (2022). Machine learning and causality in policy evaluation. Journal of Business & Economic Statistics, 40(4), 1–16. https://doi.org/10.1080/07350015.2021.1998590

[11] IZA Institute of Labor Economics. (2020). Double Machine Learning Based Program Evaluation under Unconfoundedness (IZA Discussion Paper No. 13051). IZA Institute of Labor Economics.

[12] Hays, C., & Raghavan, M. (2025). Double Machine Learning for Causal Inference under Shared-State Interference [Poster presentation]. International Conference on Machine Learning (ICML).