Automation and Machine Learning (2023)
Clausius Scientific Press, Canada

DOI: 10.23977/autml.2023.040306
ISSN 2516-5003 Vol. 4 Num. 3

The Explaimability of Double Machine Learning Causal Inference in Quasi-Natural Experiments—A Study Based on County Panel Sample Data

Zongxuan Chai[a], Tingting Zheng[*]

School of Electrical and Control Engineering, North China University of Technology, Beijing, 100144, China
[a]21101150110@mail.ncut.edu.cn
[*]Corresponding author: zhengtt2019@ncut.edu.cn

Keywords: Double machine learning; Differences-in-Differences; Machine learning explanation

Abstract: This paper focuses on the differences in the application of measurement and machine learning methods when making causal inferences. Newly highlighted double machine learning offers new research methods for policy or intervention evaluation in economic research panel data interpretability remains in doubt. We select the county panel data of Fujian Province, establish a quasi-natural experiment, and adopt a general double machine learning model and a differences-in-differences (DID) model to evaluate the policy effect of the new town policy on the optimisation of industrial structure, respectively. Both results demonstrate that the new town policy has a significant optimisation impact on industrial structure, but the dual machine learning results differ under this sample by the influence of the algorithm's advance selection. The stability test proves that the DID is more stable, and the economic significance is more explained, which is contrary to the premise of the universality of double machine learning.

1. Introduction

In the past research, the evaluation of policy effects is an important research direction for alias economics and statistics. Traditional experiments to assess policy effects have mainly used statistical methods such as Differences-in-Differences, Synthetic control, regression discontinuity design, propensity score matching, Instrumental Variables, etc. for causal inference. Shen, Y.et al.(2022)[1] Evaluating traditional models of causal analysis, which tend to be more universal and generalised at this stage. For example, differences-in-differences needs to satisfy the common trend assumption before the policy occurs with a placebo test for policy effects, and synthetic control needs to satisfy the local randomisation assumption and the continuity assumption. In addition to this, preliminary linear assumptions have been established based on the study population before conducting regression analyses, which may result in biased model assumptions, as well as vulnerability to the curse of dimensionality in the face of the selection of high-dimensional variables. Many scholars have introduced machine learning in the study of economic problems. It provides non-parametric estimation for high-dimensional data processing without model assumptions, such as using Lasso, random forest, gradient boosting, neural networks, etc. Chernozhukov.et al. (2018)[2] further proposes double machine learning(DML) with higher applicability, which corrects for the penalty of regularisation bias introduced by the direct use of machine learning.

DML havs received extensive attention from Chinese scholars since its emergence. Yang, Li.et al. (2022) [3]studied the difference between different learning model choices estimated by Monte Carlo simulation, giving different sample characteristics to choose the appropriate learning model to reduce bias. In terms of specific policy assessment, Ruting Wang, R.et al. (2022)[4] use DML while constructing K-Fold artificial counterfactual models based on cross-validation to avoid small-sample overfitting problems and introduce the placebo test as one of the robustness tests; He, J.et al. (2022)[5] introduce the mediator testing mechanism of economic research; Yan, H.et al. (2022)[6] adds synthetic double differences for balance testing; Zhang, T. and Li, J.(2022) [7]added a synthetic double difference to test for balance; and Zhang Tao et al. added instrumental variables to test for endogeneity of variables. All of the above studies use the adjusted learning model as the main robustness test method in their research. While the traditional causal analysis models have strict hypothesis testing, the results of the tests have economic analysis significance.

In summary, this paper proposes the hypothesis that the selection of a traditional econometric model for the main model of the study may be more economically interpretable, provided that the stringent assumptions are met.

2. Study Design

2.1 Study Methodology

In this paper, in order to conduct a comparative study, the general models of DML and DID are constructed respectively, and the policy effects of the pilot policy of new urbanisation on the optimisation of industrial structure are examined with reference to the panel data of counties in Fujian Province selected by 8. Jiang, A. and Yang, Z. (2020) [8] and other studies. The specific model settings are as follows:

We refer to Chernozhukov.et al.(2018)[2] to construct a general DML model:

$$t\_p_{it} = treat_{it} \times post_{it} \tag{1}$$

$$Indup_{it} = \theta_0 t\_p_{it} + L(Xlist_{it}) + U_{it}, E[U_{it}|Xlist_{it}, t\_p_{it}] = 0 \tag{2}$$

$$t\_p_{it} = M(X_{it}) + V_{it}, E[V_{it}|Xlist_{it}] = 0 \tag{3}$$

In order to calculate the estimated coefficient of policy effect, the function L of equation (2) is estimated first. Secondly, the t_p_{it} constitutes a vector to do orthogonal elimination of the influence of the control variable group, select the appropriate learning model to estimate the coefficient of disposal, and then establish an auxiliary regression to eliminate the regularity bias. Existing research are clear combing model calculation process, and its specific algorithm is not the focus of this paper, no longer too much elaboration.

Referring to Guo, C. and Zhang, W.(2018) [9]Constructing generalised DID equations:

$$Indup_{it} = \alpha + \beta(t\_p_{it}) + \gamma Xlist_{it} + \delta\mu_{it} + \vartheta\sigma_{it} + \varepsilon_{it} \tag{4}$$

In the Equation, *i* is the individual county and *t* is the year. The interaction term t_p_{it} is a multi-temporal dummy variable for the occurrence of new urbanisation policies. The $treat_{it}$ is a grouping variable, and the $postit_{it}$ is a time variable. $Indup_{it}$ is the level of industry structure, $Xlist_{it}$ is the set of control variables, and U_{it} and V_{it} are the residual terms. Among them, individual fixed and time fixed are added in DML and DID, equation (4) μ_{it} and σ_{it} are individual fixed and time fixed respectively, and ε_{it} is the error term.

2.2 Data description

The explanatory variables refer to Gan Chunhui[10] The ratio of value added of tertiary and secondary industries is chosen to measure the optimisation of industrial structure. The core explanatory variable is a dummy variable (t_p_{it}) based on the State Council's promulgation of pilot county cities of new towns. The four main control variables, namely, financing vitality (*Fin*), financial support (*Gov*), social security(*Sec*), and economic level(*Eco*), which are related to industrial structure, are selected for the study. Financing dynamism is the ratio of loans from financial institutions to residents' savings; social security is measured by the ratio of the number of resident population in the county (10,000) to the number of beds, which is obtained by summing the number of beds in hospitals and health centres and the number of beds in all kinds of social welfare adoptive units; fiscal support is the ratio of fiscal expenditure to GDP, and the level of the economy is GDP. To reduce errors caused by nonlinearity, this paper introduces the quadratic terms of the control variables into the controlled variables. The data from Fujian Provincial Bureau of Statistics, regional statistical yearbooks, etc.

3. Analysis of empirical results

In this paper, DML selection is more adapted to non-parametric random forest algorithm with large samples. In Table 1, model (1) and model (2) are the results of DID, model (3) and model (4) are the results of DML, in which the machine learning sets the sample split ratio as 1:4. Control variable2 is the quadratic term of the control variable in the table. Based on the regression results, it can be seen that the regression coefficients of the two models are significantly positive, and the difference is very small, so it can be obtained that the new town selection has a significant positive effect on the optimisation of industrial structure.

Table 1: Baseline regression results

Variable	**(1)**	**(2)**	**(3)**	**(4)**
T_p	0.165** (0.068)	0.168*** (0.064)	0.152 *** (0.064)	0.154 ** (0.065)
Control variable	Yes	Yes	Yes	Yes
Control variable2	No	Yes	No	Yes
Individual Fixed	Yes	Yes	Yes	Yes
Time Fixed	Yes	Yes	Yes	Yes

Note: T-value are in parentheses, *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

4. Robustness test

To further analyse the explanatory differences, this paper discusses the general robustness test of DID and DML with reference to the established literature. As the existing literature on the policy transmission mechanism is mainly based on methods such as constructing mediating variables and constructing DID interaction terms, scholars have debated on the research methods of the mechanism. While the solution to the research endogeneity problem are used in the instrumental variable approach or replacement policy, the selection of instrumental variables is usually not general, and are common problems faced by both. This paper does not address the question of how policy mechanisms work with endogeneity.

4.1. Parallel Trend Test and Analysis

Before use DID, the experimental group and the control group need to meet the assumption of a common trend before the policy occurs, according to the results of Figure 1 can be obtained before the policy occurs there is no significant difference, to meet the common trend, and in the second year after the initiative is issued there is a difference, which can be based on further economic significance of the analysis, there may be a policy lag effect. Based on the predictive trend results, it is observed that the regression coefficients gradually approach zero after the third year of policy implementation. This suggests a possible gradual weakening of the long-term policy effects.

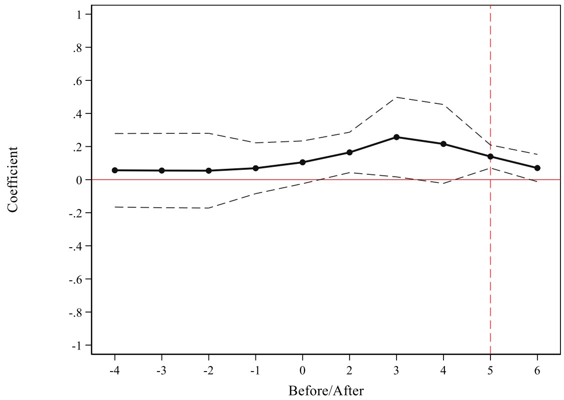

Figure 1: Parallel trend test.

4.2. Placebo Test

In order to test that the policy impact is not affected by other unobservable factors, this paper refers to the practice of existing literature to conduct a placebo test. In this paper, the experimental and control groups are randomly selected from all the samples, and then the randomised counties are merged into the original dataset that has been processed, and the randomised interaction terms are put into the regression equation for repeated regression to get the 500 estimated coefficients of the interaction terms, and the results are shown in Fig. 2, which shows that most of the coefficients are not significant and far away from the true coefficient values, and the results are still robust.

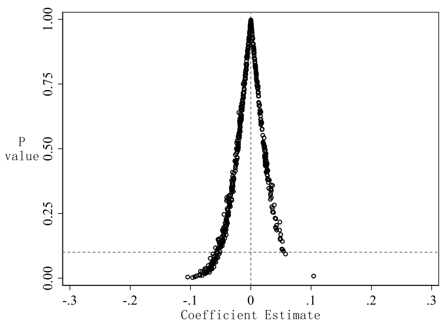

Figure 2: Placebo test results.

4.3. Propensity Matching

To avoid systematic differences between the experimental and control groups of the study subjects. In this study, propensity score matching was adopted for testing. The propensity score values are obtained after logit regression of the policy's dummy variables on the control variables, and the specific results are shown in Table 2.

It can be found that the p-values of the matched covariates become insignificant except for finance, and there is no significant systematic difference between the experimental and control groups. At the same time, the standard deviation of all covariates decreased significantly. The standard deviation is not less than 10% may lie in the fact that the difference between DID and DML is not significantly analysed, the data are all unstandardised using the same group of data, and so on. The economic significance of the absence of systematic differences in their county economic and social data can be further analysed.

Table 2: Propensity to match results

	Before Matched			**After Matched**		
V	**bias (%)**	**T value**	**P value**	**bias (%)**	**T value**	**P value**
Gov	-65.6	-3.25	0.001	-12.3	-0.59	0.558
Soc	-46.4	-2.47	0.014	-24.5	-1.19	0.237
Fin	-18.4	-0.91	0.364	10.2	0.60	0.554
Eco	84.6	8.91	0.000	-2.4	-0.13	0.894

4.4. Adjust the machine learning model

Table 3: Results after Adjust the machine learning model

Variable	**Adjust the cross-fitting split ratio**		**Replacement of machine learning models**			
	1:2	**1:7**	**Lassocv**	**Lassocv**	**Neural networks**	**Neural networks**
T_p	0.194*** (0.069)	0.168*** (0.064)	0.199*** (0.069)	0.194*** (0.069)	-69.762*** (23.315)	1.237*** (0.016)
Control variable	Yes	Yes	Yes	Yes	Yes	Yes
Control variable2	Yes	Yes	No	Yes	No	Yes
Individual Fixed	Yes	Yes	Yes	Yes	Yes	Yes
Time Fixed	Yes	Yes	Yes	Yes	Yes	Yes

Note: T-value are in parentheses, *, **, and *** indicate significance at the 10%, 5%, and 1% levels, respectively.

In this paper, we refer to the general literature robustness test to adjust the cross-fitting split ratio of machine learning to 1:2, 1:7, with the replacement of the machine learning model for regression. Table 3 presents the results after adjusting the machine learning model, indicating minor coefficient changes due to adjustments in the segmentation ratio. However, upon substituting the machine learning model, there are significant differences in the regression coefficient results. Specifically, when set as a neural network model without incorporating quadratic controlled variables, the coefficients of the neural network model show small negative values, contrary to the baseline regression in this paper. This demonstrates that the settings of machine learning models influence regression results, suggesting that pre-experiment assumptions of machine learning models might lead to biases in regression outcomes.At the same time, the test does not have the significance of further economic analysis.

5. Conclusions

This paper investigates the discussion of the interpretability of DML in panel sample data. In the comparison of DML with the traditional causal analysis method of DID, it is found that the two in the benchmark regression under the premise assumption of satisfying the parallel trend as an example, there is no big difference between the regression results of DML and DID. However, in the DML learning model algorithm selection there is a replacement algorithm regression coefficient difference in the results, so DML may exist model algorithm advance assumptions brought about by the results of bias, need to choose and data characteristics matching algorithm.

Therefore, combined with the results of the analysis of this paper, under more stringent assumptions, the results of DID are more robust and economically interpretable, based on the new urbanisation policy studied in this paper, the curse of dimensionality and linear assumption bias eliminated by dual machine learning has less impact, and there is no significant difference between the two models, but the results of the double differencing are more robust, and dual machine learning is more suitable for auxiliary testing, and the assumptions of this paper are Consistent. In economic research, the DID model, for example, its need to meet the hypothesis test of the same economic analysis, such as parallel trends can be inferred from the policy effect of time trend changes, while the simple machine learning method lacks more significance of the solution, need to carry out other model processing to reflect. For the applicability of traditional causal analysis methods and the development of artificial intelligence under the new machine learning methods in this paper, the starting point is still more basic, but also many scholars to continue in-depth analysis.

Acknowledgement

Innovation and Entrepreneurship Training Program for College Students at North China University of Technology (X202310009117).

References

[1] Shen, Y.; Li, X.; Zhou, Q. Estimation and Inference of Treatment Effects with Panel Data in the Big Data Era. Quantitative & Technical Economics 2022, 39, 120.

[2] Chernozhukov, V; Chetverikov, D; Demirer, M; Duflo, E; Hansen, C; Newey, W; Robins, J. Double/debiased machine learning for treatment and structural parameters. ECONOMETRICS JOURNAL 2018, 21, C1-C68.

[3] Yang, L.; Zhao, J.; Li, Q. Monte Carlo simulation of Treatment Effect Estimation of Double Aachine learning. Statistics and Decision 2022, 6, 26-31.

[4] Wang, R.; Peng, F.; Li, W. Does Terminating Rigid Payment Diminish Financing Cost of Companies? Management World 2022, 38(04), 42-64.

[5] He, J.; Peng, F.; Xie, X. Mixed-ownership Reform, Political Connection and Enterprise Innovation: Based on the Double/Unbiased Machine Learning Method. Science and Technology Management Research 2022, 42(11), 116-126.

[6] Yan, H.; Wang, J.; Sun, J. How Does the Entry of Collective Construction Land into the Market Affect the State-Owned Construction Land Market. Quantitative & Technical Economics 2023, 40(6), 195-216.

[7] Zhang, T.; Li, J. Network Infrastructure,Inclusive Green Growth, and Regional Inequality: From Causal Inference Based on Double Machine Learning. Quantitative & Technical Economics 2023, 40(4), 113-135.

[8] Jiang, A.; Yang, Z. New Urbanization Construction and High—quality Urban Economic Growth: an Empirical Analysis Based on the Double Difference Method. Inquiryin to Economic Issues 2020, 3, 84-99.

[9] Guo, C.; Zhang, W. The impact of new-type urbanization construction on the quality of regional economic development under the background of industrial structure upgrading: empirical evidence based on PSM-DID. Industrial Economics Research 2018, 5, 78-88.

[10] Gan, C.; Zheng, R.; Yu, D. Cyclical Fluctuations and Nonlinear Dynamics of Inflation Rate. Economic Resrarch Journal 2011, 46(5), 4-16+31.