Title and Information

Machine Learning Application on DID Estimation

Oscar Lu, Faye Yang, Linfeng Shi, Qijin Liu, Weizi He
Texas A&M University
April 24, 2024

Problem Statement

Introduction This paper investigates the effects of minimum wage increases on low-wage jobs using a difference-in-differences design. The analysis focuses on 138 state-level changes from 1979 to 2016, examining employment shifts around the new wage minimums. Results indicate that overall employment levels in low-wage jobs remain stable, with significant wage spillovers at the lower end of the wage distribution.

Description This study employs a multi-period DID analysis complemented by a double lasso technique to examine the effects of minimum wage changes on employment across U.S. states from 1993 to 2003.

Objective To enhance the parallel assumption before treatment in our DiD regression for a better model fit to data.

Methodology

Modeling Technique: Leveraged the Difference-in-Differences (DID) framework, coupled with machine learning algorithms, to meticulously select control groups, thereby enhancing the precision of the analysis.

Causal Inference Analysis: Utilized advanced machine learning methodologies to refine policy effect estimation, fostering a deeper understanding of causal relationships within the data.

Data Preprocessing: Implemented rigorous data preprocessing techniques to establish consistency and comparability across groups, ensuring the reliability and validity of the subsequent analysis.

Analysis

Twoway Approach: Using Causal Forest to generate heterogenous effects, focusing on weighted averages for various subpopulations.

Main Approach: Double Lasso for enhanced control group selection to match Arkansas as a comparative state, solidifying the parallel assumption in DID modeling.

ML Enhancement Summary

The double lasso technique serves as a sophisticated statistical method to refine the selection of a control group for our Difference-in-Differences (DID) analysis. By implementing this approach, we were able to rigorously filter through a pool of potential controls and identify a state—Arkansas—whose pre-treatment characteristics and trends closely mirrored those of the treatment group. This careful matching process is pivotal in reinforcing the parallel trends assumption, a foundational requirement for the validity of DID estimates.

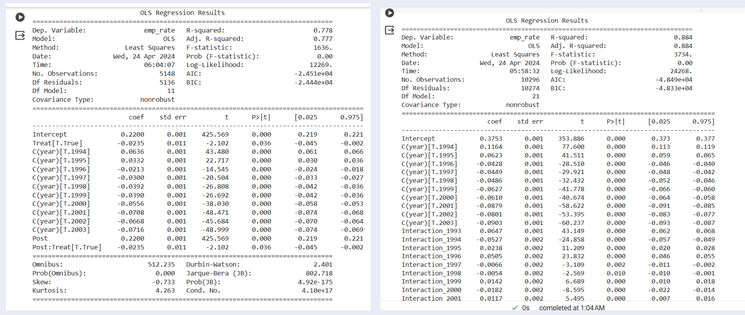
This assumption posits that, in the absence of the treatment—in this case, the policy change under study—the treatment and control groups would have followed the same trajectory over time. Ensuring this assumption holds true allows us to more accurately attribute any post-treatment differences in outcomes to the intervention itself, rather than to pre-existing or unrelated trends. Through the application of the double lasso method, our study was able to strengthen the parallel assumption by effectively accounting for a multitude of observed and unobserved covariates that could influence the employment rates we sought to examine. By doing so, we enhanced the robustness of our DID model, bolstering our confidence in the conclusions drawn about the policy's impact.

Findings

Correlation: Significant relationship between worker numbers and minimum wage.

State Adjustments: Diverse impacts of state-level minimum wage changes.

Summary: In conclusion, we selected Arkansas as an example based on the percentage change in MW of the treatment group and used Double LASSO to filter out New York as a control group, and we found significant results that strengthened the conclusions of the original study.



OLS Regression Results

Dep. Variable:	emp_rate	R-squared:	0.778
Model:	OLS	Adj. R-squared:	0.777
Method:	Least Squares	F-statistic:	1636.
Date:	Wed, 24 Apr 2024	Prob (F-statistic):	0.00
Time:	06:04:07	Log-Likelihood:	12269.
No. Observations:	5148	AIC:	-2.451e+04
Df Residuals:	5136	BIC:	-2.444e+04
Df Model:	11		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.2200	0.001	425.569	0.000	0.219	0.221
Treat[T.True]	-0.0235	0.011	-2.102	0.036	-0.045	-0.002
C(year)[T.1994]	0.0636	0.001	43.480	0.000	0.061	0.066
C(year)[T.1995]	0.0332	0.001	22.717	0.000	0.030	0.036
C(year)[T.1996]	-0.0213	0.001	-14.545	0.000	-0.024	-0.018
C(year)[T.1997]	-0.0300	0.001	-20.504	0.000	-0.033	-0.027
C(year)[T.1998]	-0.0392	0.001	-26.808	0.000	-0.042	-0.036
C(year)[T.1999]	-0.0390	0.001	-26.692	0.000	-0.042	-0.036
C(year)[T.2000]	-0.0556	0.001	-38.030	0.000	-0.058	-0.053
C(year)[T.2001]	-0.0708	0.001	-48.471	0.000	-0.074	-0.068
C(year)[T.2002]	-0.0668	0.001	-45.684	0.000	-0.070	-0.064
C(year)[T.2003]	-0.0716	0.001	-48.999	0.000	-0.074	-0.069
Post	0.2200	0.001	425.569	0.000	0.219	0.221
Post:Treat[T.True]	-0.0235	0.011	-2.102	0.036	-0.045	-0.002

Omnibus:	512.235	Durbin-Watson:	2.401
Prob(Omnibus):	0.000	Jarque-Bera (JB):	802.718
Skew:	-0.733	Prob(JB):	4.92e-175
Kurtosis:	4.263	Cond. No.	4.10e+17

OLS Regression Results

Dep. Variable:	emp_rate	R-squared:	0.884
Model:	OLS	Adj. R-squared:	0.884
Method:	Least Squares	F-statistic:	3734.
Date:	Wed, 24 Apr 2024	Prob (F-statistic):	0.00
Time:	05:58:32	Log-Likelihood:	24268.
No. Observations:	10296	AIC:	-4.849e+04
Df Residuals:	10274	BIC:	-4.833e+04
Df Model:	21		
Covariance Type:	nonrobust		

	coef	std err	t	P>|t|	[0.025	0.975]
Intercept	0.3753	0.001	353.886	0.000	0.373	0.377
C(year)[T.1994]	0.1164	0.001	77.600	0.000	0.113	0.119
C(year)[T.1995]	0.0623	0.001	41.511	0.000	0.059	0.065
C(year)[T.1996]	-0.0428	0.001	-28.510	0.000	-0.046	-0.040
C(year)[T.1997]	-0.0449	0.001	-29.921	0.000	-0.048	-0.042
C(year)[T.1998]	-0.0486	0.001	-32.432	0.000	-0.052	-0.046
C(year)[T.1999]	-0.0627	0.001	-41.778	0.000	-0.066	-0.060
C(year)[T.2000]	-0.0610	0.001	-40.674	0.000	-0.064	-0.058
C(year)[T.2001]	-0.0879	0.001	-58.622	0.000	-0.091	-0.085
C(year)[T.2002]	-0.0801	0.001	-53.395	0.000	-0.083	-0.077
C(year)[T.2003]	-0.0903	0.001	-60.237	0.000	-0.093	-0.087
Interaction_1993	0.0647	0.001	43.149	0.000	0.062	0.068
Interaction_1994	-0.0527	0.002	-24.858	0.000	-0.057	-0.049
Interaction_1995	0.0238	0.002	11.209	0.000	0.020	0.028
Interaction_1996	0.0505	0.002	23.832	0.000	0.046	0.055
Interaction_1997	-0.0066	0.002	-3.109	0.002	-0.011	-0.002
Interaction_1998	-0.0054	0.002	-2.569	0.010	-0.010	-0.001
Interaction_1999	0.0142	0.002	6.689	0.000	0.010	0.018
Interaction_2000	-0.0182	0.002	-8.595	0.000	-0.022	-0.014
Interaction_2001	0.0117	0.002	5.495	0.000	0.007	0.016

✓ 0s completed at 1:04 AM

Figure: Regression results before and after optimization.

Future Research

- Assess long-term model accuracy: Conduct comprehensive evaluations over extended time periods to ascertain the durability and reliability of the model's predictive capabilities.
- Validate research outcomes by examining diverse demographic groups and economic sectors, ensuring the robustness and generalizability of the findings.
- Expand the scope of analysis by incorporating additional variables into the model, facilitating a more comprehensive understanding of the underlying mechanisms driving the observed phenomena.