Evaluating the Effects of Financial Deregulation on Bank Risk using Double Machine Learning

Gaurav K. Shah

Thesis submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
in
Computer Science and Applications

Sara Hooshangi, Chair
Ali Habibnia, Co-Chair
Chang-Tien Lu, Co-Chair

May 12, 2025
Blacksburg, Virginia

Keywords: Some Keywords, Subject matter, etc.

Evaluating the Effects of Financial Deregulation on Bank Risk using Double Machine Learning

Gaurav K. Shah

(ABSTRACT)

This work examines the causal impact of deregulation within the U.S. banking sector, focusing on the rollback of a specific provision of the Dodd-Frank Act through the Economic Growth, Regulatory Relief, and Consumer Protection Act of 2018 (EGRRCPA). Originally enacted in response to the 2008 financial crisis, the Dodd-Frank Act introduced extensive regulatory reforms aimed at mitigating systemic risk. However, the partial repeal of its provisions has prompted renewed interest in assessing the implications for bank risk and financial stability. Our research contributes to this growing body of work by employing recent developments in causal inference, particularly Double Machine Learning (DML), to more accurately estimate treatment effects. DML leverages machine learning algorithms to flexibly model both treatment and outcome processes, controlling for bias via orthogonalization techniques and sample-splitting strategies. By applying DML to panel data, we address the complexities of policy evaluation with panel data and aim to improve the robustness of causal estimates. We conduct a reanalysis of Chronopoulos et al. [12], comparing estimates produced using traditional fixed effects with linear regression models and those generated by modern machine learning based estimators. Furthermore, we investigate the implications of key implementation choices such as panel data transformation techniques, cross-fitting procedures, and hyperparameter optimization on the performance and interpretability of DML

in applied policy settings. Our work underscores the value of integrating modern computational tools into empirical regulatory analysis, offering insights for policymakers and causal researchers.

Evaluating the Effects of Financial Deregulation on Bank Risk using Double Machine Learning

Gaurav K. Shah

(GENERAL AUDIENCE ABSTRACT)

This research explores the impact of the 2018 rollback of a key Dodd-Frank provision on bank risk in the U.S. financial sector. We combine tools from economics and computer science, specifically causal inference and machine learning, to estimate how these policy changes influenced bank behavior. Using Double Machine Learning (DML), we improve upon traditional econometric approaches by accounting for complex relationships and potential confounders in the data. Our results show that machine learning methods can provide accurate and nuanced estimates of policy effects, offering valuable insights for both researchers and regulators interested in data-driven financial policy.

Dedication

To my friends and family

– for their unwavering support. Thank you.

Acknowledgments

I would like to express my sincere gratitude to Ali Habibnia for his mentorship, generosity, and support throughout this process.

I would also like to thank Muhammed Yilmaz for sharing the data that made this research possible, and to Annalivia Polselli, Dominik Papies, and Jonathan Fuhr for their time and insights.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In recent years, there has been a growing interest in causal machine learning methods, driven by the widespread application of machine learning techniques, due to their strong predictive power, and the pursuit of uncovering causal relationships in data. The body of work on causal inference has a long history in the applied social sciences and economics, largely because of the need to distinguish correlation from causation in observational data, where randomized controlled trials are often infeasible or unethical. This is particularly important in policy analysis, where understanding the causal impact of interventions, such as education programs, tax reforms, or other macroeconomic regulations, is essential for informed decision making.

One method in particular that has become increasingly popular is Double Machine Learning (DML), also known as Double/Debiased Machine Learning or Orthogonal Machine Learning. This technique uses machine learning to estimate the effect that covariates in a dataset have on both the outcome and the treatment. These estimates, referred to as nuisance parameters, are then used within a framework that applies the concept of Neyman orthogonality to define moment conditions. These conditions are constructed to be insensitive to small errors in the nuisance estimates, making the final treatment effect estimator more robust. The method aims to identify the causal effect of the treatment or intervention on the outcome, under the key assumption of unconfoundedness, which states that the treatment assignment is independent of potential outcomes, conditional on observed covariates.

Our analysis focuses on the Dodd-Frank Act which was passed in 2010 as a response to the 2008 financial crisis and sought to impose stricter regulations on banks. In 2018, certain provisions of the act were rolled back with the passage of the Economic Growth, Regulatory Relief, and Consumer Protection Act (EGRRCPA). Through a reanalysis of Chronopoulos et al. [12] we investigate the causal impact of deregulation on bank risk within the U.S. banking sector.

Our contributions are two-fold. First, we contribute to the financial regulatory space by conducting a reanalysis of Chronopoulos et al. [12]. Most works in the realm of financial policy analysis rely on linear regression-based methods of causal inference. These typically include difference-in-differences and two-way fixed effects (TWFE). The difference-in-differences method estimates the causal effect of a treatment by comparing the change in outcomes for treated and control groups before and after the treatment, under the assumption of parallel trends. TWFE extends this approach within a regression framework by including unit and time fixed effects to control for differences across units or over time that are not directly measured in the data but still influence the outcome, known as unobserved heterogeneity. The majority of literature evaluating the Dodd-Frank Act continues to employ these traditional linear methods. We aim to contribute to the small but growing body of work that seeks to demonstrate the practicality and feasibility of DML not only in financial policy research but in the more complicated panel data setting.

Second, we aim to show the feasibility of machine learning methods for causal inference on panel data, which consists of observations on multiple units (such as firms or regions) over multiple time periods. To date, there have been few empirical analyses of DML using panel data. The papers that do utilize it do not specify adjustments made to the data or method of estimation. We use recent works by Clarke and Polselli [13] and Fuhr and Papies [17] to adapt the data transformation and estimation strategies for the panel setting. Unlike

many previous applied works, we incorporate new methods of more appropriately applying DML for panel data, by adjusting the sample splitting procedure for training and testing machine learning models and incorporating fixed effects to capture heterogeneity. Following other applied works, we compare the magnitude and significance of estimates obtained via machine learning and those obtained using linear methods. Additionally, we evaluate the sensitivity of DML estimates across different hyperparameter choices. Based on our findings, we uphold the original results of Chronopoulos et al. [12], and show why some methods of performing DML are more appropriate than others.

Having been inspired by works that evaluate bank risk and regulatory burden in the Dodd-Frank era such as Powell [28] and Chronopoulos et al. [12], we seek to apply a DML framework based on Chernozhukov et al. [10] and address the challenges associated with panel data based on works by Clarke and Polselli [13] and Fuhr and Papies [17].

Chapter 2

Review of Literature

Numerous papers in the current literature attempt to establish causal links between the revision of the Dodd-Frank Act through the Economic Growth, Regulatory Relief, and Consumer Protection Act of 2018 (EGRRCPA). These papers primarily use methods such as fixed effects regression and difference-in-differences, due to the panel data setting. Furthermore, among the vast literature that surrounds policy and regulation analysis in the banking sector, many of the current methodologies use traditional causal methods that pose strong assumptions between the treatment and covariates. There is an increasing body of literature that aims to use machine learning methods along with more sophisticated causal methods, such as DML to establish stronger causal links in the realm of policy analysis, especially focusing on high dimensional data with confounding variables. Although the EGRRCPA includes several provisions that roll back regulations related to stress test frequency, leverage ratios, the Volcker Rule, examination cycles, and more, our research focuses specifically on the impact of raising the threshold for enhanced standards for bank holding companies (hereafter referred to as banks) from $50 billion to $250 billion in total consolidated assets. We perform a reanalysis of work done by Chronopoulos et al. [12] and evaluate the results of DML against their baseline finding based on a linear TWFE regression.

2.1 Related Works - Treatment Effect Analysis of Regulations in the Banking Sector

Much of the current research on financial policy analysis uses traditional methods of causal analysis. Some even run simple linear regressions as a way to demonstrate a correlation but fail to show strong causal evidence.

Powell [28] and Chronopoulos et al. [12] both use TWFE to analyze banks before and after the passage of the EGRRCPA. Powell [28] analyzes banks with less than $10 billion in total consolidated assets (known as "community banks") to see whether the Community Bank Leverage Ratio and Volker Rule revisions of the EGRRCPA, which reduced capital requirements, provided more time between examinations, and allowed for fewer reporting requirements, had any effect on regulatory burden. Chronopoulos et al. [12] uses a difference-in-differences approach to measure the change in bank risk after the EGRRCPA. Using data collected from FRY-9C forms, they measure the change in risk weighted assets from 2015 to 2020. The treatment group in this study are banks with total assets between $50 billion to $250 billion. Their findings show a significant increase in risk after passage of the EGRRCPA in 2018. Both studies use panel data.

Kim and Katchova [22] used ordinary least squares to evaluate whether stricter lending regulations by way of increased capital requirements of Basel III could restrict credit availability for farmers. They use FDIC panel data from 2008 to 2017 to examine the agricultural loan volume and growth rates for agricultural banks and all US banks. They find that agriculture loan growth rates have slowed but the amount of agricultural loan volume issuance is still positive.

Similar to Chronopoulos et al. [12], both Janda and Kravtsov [20] and Balasubramanyan

et al. [6] measure the change in bank risk following a significant policy change. Janda and Kravtsov [20] assess how regulatory stress tests affect investment decisions and portfolio choices of banks in the European Union. The primary aim is to explore the heterogeneity in the treatment groups: at the group level (comparing the stress-tested with non-stress-tested banks), at the individual unit level (the strength of the capital requirement effect for a single bank), and variation in the timing of treatment (event study). Using difference-in-differences and instrumental variable regression, they observed a decline in the riskiness of portfolios in response to the stress tests. Balasubramanyan et al. [6] analyze the impact of having a risk committee and chief risk officer on bank risk after the passage of the Dodd-Frank Act. Using a regression discontinuity and difference-in-differences based approach, they do not find a significant causal link. Bao et al. [7] uses a similar regression-based approach to investigate the liquidity of cooperate bonds based on stress events such as the passage of the Volcker rule, a key piece of regulation imposed on the banking sector after the 2008 financial crisis.

2.2 Related Works - Methodology of Similar Applied Works

We find numerous papers that use DML on cross-sectional data for stronger causal analysis especially in the context of high-dimensionality, nonlinear data, and regularization bias to avoid the traditional assumptions laid out in most causal methods. Knaus [23], Hansen and Siggaard [19], Chin et al. [11], Gao et al. [18], Yuan and Liu [31] Ellickson et al. [16], Wang et al. [29], and Yang et al. [30] all use DML to inform stronger causal estimates, citing some or all of the aforementioned benefits.

Knaus [23] uses DML to estimate the effects of musical practice on a child's cognitive and non-cognitive skills. They assess the sensitivity of estimates and parameter choices and find significant positive effects between musical practice and cognitive skills, but caution the impact that parameter choices might have on estimation. Hansen and Siggaard [19] discusses the trend of a post earnings announcement drift, where a stock's price drifts in the direction of an earnings surprise after the announcement. They use DML for variable selection and attribute overcoming omitted variable bias as one of the benefits of DML. Chin et al. [11] estimates the change in case dismissal rates under judges running opposed or unopposed in the following primary elections. The author uses DML to determine the effect of the proximity of a judicial election to judges' dismissal rates. This paper used DML with lasso, decision trees, and random forests. The data consists of individual records of cases in Pennsylvania. The results show that the difference in dismissal rates is constant for judges in contested and uncontested elections, up to 6 months before a primary election, when dismissal rates start to fall. Ellickson et al. [16] utilizes DML to analyze the heterogeneity in marketing emails to identify the causal estimates for specific characteristics of the email, primarily related to the subject line of the email. The results show statistically significant effects of emails on engagement levels. The authors discuss that unbinding the separate impact of the different components of an ad campaign is challenging due to the high dimensionality, selection bias, and low statistical power. DML is able to better capture the heterogeneity of the treatment components in the subject line and the heterogeneous customer engagement. Wang et al. [29] use a general DML and propose a robust DML procedure that uses median machine learning methods instead of traditional machine learning methods to analyze the impact of cerebrospinal fluid on Alzheimer's disease severity. Yang et al. [30] analyze the impact of Big N auditors (professional audit firms) on audit quality.

To address its applications in a broader econometric context, Baiardi and Naghi [5] use

DML with causal forest and other generic machine learning methods to reanalyze Djankov et al. [15], Nunn and Trefler [27], DellaVigna and Kaplan [14], and Loyalka et al. [26]. The authors seek to evaluate the performance of causal machine learning versus traditional machine learning methods when the relationship between outcome, covariates, and treatment is nonlinear. They show that when the number of covariates increases relative to the sample size, DML outperforms ordinary least squares in both the linear and nonlinear case. They reevaluate 2 studies for the average treatment effect and 2 for the heterogeneous treatment effect. For the average treatment effect, the authors reevaluate Djankov et al. [15] which analyzes the effect of corporate taxes on investments. Using DML they find the coefficient is greater in magnitude and obtain lower standard errors. The authors reevaluate Nunn and Trefler [27] which investigates the effect of skill-biased tariffs on growth. They use DML to estimate the average treatment effect and obtain coefficients that are much smaller in magnitude and nonsignificant compared to the original paper, suggesting that the correlation between skill-biased tariffs and long-term economic growth is not robust. Regarding the heterogeneous treatment effect, they reevaluate DellaVigna and Kaplan [14] which aims to measure the effect of Fox News viewership on Republican vote share. The original authors showed that towns where Fox News became available between 1996 and 2000 had an increase in the republican vote share. Using causal forest as the machine learning model, they upheld results from the original paper. Finally, they reevaluate the effect of teacher training on student performance, as in Loyalka et al. [26], and find no significant impact of professional development on student outcomes.

Traditionally in economics, to avoid omitted variable bias, regressions aimed at identifying a causal relationship would include many controls. Traditional machine learning methods optimize the goodness of fit for predictive analysis rather than identifying the causal effect. In empirical economic research, if standard machine learning techniques are used to estimate

a causal effect this will result in biased estimates. DML aims to overcome this challenge by using standard machine learning techniques for causal inference. Baiardi and Naghi [5] establish 4 main reasons for the use of DML. Firstly, DML estimates are more robust to potential nonlinear confounders. Secondly, causal machine learning methods are better suited for large numbers of covariates as they assume a sparse model and use regularized regressions. Thirdly, it allows for model selections while ensuring that relevant transformations are considered. Lastly, due to the implementation of machine learning, it is helpful in estimating heterogeneous treatment effects, compared to manually modeling different interaction terms.

Kumar et al. [24] investigate the returns of actively and passively managed funds under the effect of interest rate changes by the U.S. Federal Reserve from January 1986 to December 2021. They analyze gradient boosted and linear regression models using the DML approach and find that a 1 percent interest rate increase causes and actively managed fund's returns to decrease by almost 12 percent. Similar to other papers on the topic of DML, the authors highlight its advantage in handling high-dimensional data and addressing confounding. They further mention the gradient boosting is beneficial in handling non-linearity. They use the XGBoost library in Python and the DoubleML library to employ the DML method. The authors mention that future research is needed to investigate the assumptions of the methodology.

Baiardi and Naghi [4] use DML to conduct a replication of Alesina et al. [1], which investigates the effect of plough agriculture on gender roles. The original study hypothesizes that cultures which traditionally used ploughs for agriculture have a stronger gender division of labor, as men had the physical advantage in plough cultivation, and that this division persists in those societies today. Their findings support Alesina et al. [1], as they find a significant negative treatment effect. In their replication, Baiardi and Naghi [4] find an even stronger effect that they attribute to the relaxed assumptions and flexibility of being able to capture more

controls with DML.

Li et al. [25] investigate the psychological links between depression and suicide among users of a social media "depression community". They use DML to analyze linguistic features extracted from Weibo (a Chinese social media app) posts and investigate the pathways linking depression to suicide risk. The authors find that depression, as characterized by the treatment group, has a significantly higher linkage to suicide risk than the control group. Traditional methods like ordinary least squares face challenges in detecting the heterogeneous effects of depression on suicide risk and the interactions between multiple features. This limitation highlights a common argument for the use of DML, as it allows for the identification of both linear and nonlinear trends.

To date, few studies have properly used DML in a panel data setting and those that have, do not explicitly state how they captured the fixed effects necessary to properly model the unobserved heterogeneity in panel data. Chai et al. [8] uses panel data to evaluate the explainability of DML versus difference-in-differences. They find that after conducting stability tests, difference-in-differences is more stable than DML and the economic significance is better explained. They argue that this goes against the universality of DML. They use a random forest algorithm as the underlying learner and a 1:4 sample split to train the data. The authors find that, under the parallel trends assumption, both the traditional difference-in-differences approach and DML with individual and time-varying fixed effects show a significant treatment effect. The difference between the coefficient obtained via the difference-in-differences and DML coefficient is very small. Both Yuan and Liu [31] and Gao et al. [18] follow a similar approach by applying the traditional DML framework on panel data. They examine policy implications on Chinese companies. Yuan and Liu [31] investigate the Made in China 2025 initiative on urban economic growth. They analyze panel data of several hundred Chinese cities from 2006 to 2021, of which 30 are a part of the treatment

group as pilot cities (cities that adopted the Made in China 2025 policy). They use lasso along with DML to conduct a baseline regression that included fixed effects for the city and time. Gao et al. [18] assess a company's green technological innovation behavior based on whether the firm is located in a pilot city for a new industrial land use policy. Jiang et al. [21] similarly uses DML on panel data to analyze the effect of the digital economy on urban ecological development using data from Chinese cities. Kumar et al. [24] also includes panel data in their analysis. Although these papers use DML on panel data, they do not formally specify the transformations or alterations made to the data to account for the fixed effects.

There are an increasing number of works that seek to adapt DML for different settings. Chang [9] develops a score function that satisfies the Neyman orthogonality condition to calculate the treatment effect of a difference-in-differences design. They compare the DML estimates to semi-parametric difference-in-difference estimates and finds the DML estimates to be more stable in converging to the true causal parameter, through several Monte Carlo simulations. However, Chang [9] mainly focuses on cross sectional data and does not implicitly account fixed effects that we would need to address in a panel setting. Fuhr and Papies [17] and Clarke and Polselli [13] are recent studies that evaluate DML and adapt it specifically for panel data. The authors explore methods for incorporating panel data into DML, comparing various data transformations and cross-fitting strategies. Clarke and Polselli [13] proposes a cross-validation scheme that ensures all time-varying observations for a given unit are included within the same fold to account for the potential time dependence. Both studies examine the effects of different data transformations on panel data across multiple data-generating processes. In a partially linear setting, Fuhr and Papies [17] find that DML with correlated random effects yields the best performance in terms of convergence to the true causal parameter across various Monte Carlo simulations, with different relationships between the treatment, response, and covariates. Conversely, Clarke and Polselli [13] find

that the first-differencing method produces the most robust estimates under the partially linear setting, particularly when using time-based sample splitting, where the entire time-series of a sample unit is contained within the same fold.

Chapter 3

Background and Data

3.1 Background

In response to the 2008 financial crisis, the United States Congress passed the Dodd-Frank Wall Street Reform and Consumer Protection Act, commonly known as the Dodd-Frank Act. This landmark legislation comprised of a series of regulations aimed to enhance regulatory oversight, reduce systemic risk, and protect consumers within the financial sector. Key provisions of the bill included the establishment of the Consumer Financial Protection Bureau, which imposed stricter transparency requirements on mortgage and consumer lenders, and the implementation of the Volker rule, which curtailed speculative trading activities and prohibited banks from engaging in proprietary trading. It also introduced enhanced oversight of credit default swaps and derivatives, financial securities that played a central role in the 2008 crisis.

Approximately a decade later, the EGRRCPA of 2018 was signed into law, amending several regulations of Dodd-Frank. These revisions, enacted between 2018 and 2019, were designed to reduce regulatory burdens, particularly on smaller banks. Some key changes included, eliminating the Volker rule provision for banks with under $10 billion in assets, raising the threshold for capital and liquidity standards, and eliminating mandatory company-run stress tests for certain banks. One notable provision, Tailoring Capital and Liquidity Rules for Large Banking Organizations, increased the threshold for more stringent capital and liquidity

requirements from banks with $50 billion to $250 billion in assets under management. This study focuses on the effects of that specific revision. Since banks that were previously treated under this provision (i.e., those with between $50 and $250 billion in assets) became exempt of strict capital and liquidity requirements at the time of the rule change, it is reasonable to hypothesize that they subsequently took on increased risk following the deregulation. We conduct a reanalysis of the findings reported in Chronopoulos et al. [12] using DML, an advanced causal framework that integrates machine learning with traditional econometric techniques. By leveraging the flexibility of DML, we aim to generate more robust and accurate estimates of the policy's causal impact on bank risk.

3.2 Data

An important aspect of this research is the use of panel data in the causal setting. This type of data is defined as tracking the same units (banks) over multiple time periods (quarters of a year). This can be viewed as a combination of cross-sectional data (observations of different units at one time period) and time series data (observations of a single unit over multiple time periods). The data originally comes from the National Information Center of the Federal Financial Institutions Examination Council. The data is collected from FR Y-9C files, which are reports containing financial data of all domestic bank holding companies. The data is collected quarterly and includes thousands of banks identified by their unique ID, known as an *RSSD ID*. The data contains financial information on each bank such as *Total Consolidated Assets*, *Interest Bearing Deposits*, *Collateralized Loan Obligations*, *Amount of US Treasuries*, *Risk Weighted Assets*, and much more. We use the same covariates from Chronopoulos et al. [12], they include: *Deposit Funding*, *Provisions Ratio*, *Operating Efficiency*, *Dividend Payout Ratio*, *Derivatives Ratio*, and *Liquidity Ratio*. These six variables are chosen as controls due

to their popularity in empirical investigations of bank risk. We summarize their importance according to Chronopoulos et al. [12] below.

Deposit Funding is calculated by dividing the interest bearing deposits by total assets. This shows the extent to which a bank relies on customer deposits as a source of funding. This variable is generally considered to provide a stable source of funds, which may contribute to the overall financial stability of the institution. Conflicting literature shows that banks with a higher proportion of deposits have stronger incentives to avoid engaging in excessively risky activities in order to protect their charter value. Conversely, banks that rely heavily on retail deposits may be inclined to take more risk as the presence of deposit insurance acts as a safety net.

Provisions Ratio is the ratio of loan loss provisions to total assets. This variable measures a bank's approach to managing credit risk and helps them to engage in earnings management as a way to show smooth profits across periods. However, excessive use of provisioning may contribute to greater complexity and opacity, which can impede oversight from regulators. The potential for higher complexity and reduced transparency makes it more challenging to assess the true risk exposure.

Operating Efficiency is calculated as the ratio of non-interest expenses to the sum of non-interest income and net interest income. This variable helps to reflect the organizational and administrative effectiveness of the bank. Higher values are associated with excessive overhead costs and indicate lower efficiency which are expected to increase risk.

Dividends are included as the ratio of total common stock dividends to total assets. Dividend policies offer insight into a bank's risk appetite and financial signaling strategies. This variable also has conflicting explanations about its effect on risk. Higher dividends may be associated with greater risk when associated with risk-shifting behavior, such as management

transferring wealth to shareholders. In contrast, consistent dividend payouts can serve as a positive signal to the market by indicating confidence in the bank's financial health and reducing perceived risk.

Derivatives also play a complex role in perceived risk. When used for hedging, derivatives can be effective tools for mitigating specific types of financial risk. In contrast, they can also be used to build leverage and serve as an indication of speculative trading thereby increasing risk.

Liquidity Ratio is the ratio of cash and cash equivalent holdings to total assets. This is a critical measure of the health of a bank as it reflects the ability for a bank to meet immediate or unforeseen demands for cash. This could be especially important in times of financial stress. Institutions that fund long-term assets with short-term liabilities expose the bank to liquidity risk, which can lead to destabilizing bank runs. Holding higher proportions of liquid assets reduces risk and makes the institution more resilient to shocks. Post-financial crisis regulatory frameworks, such as Basel III, have emphasized liquidity management as a central component of supervision.

The central outcome that we aim to observe is the log-change in *Risk Weighted Assets*. The balanced panel data is only available after quarter 2 of 2015. Our observations range from quarter 2 of 2015 to quarter 1 of 2020. Chronopoulos et al. [12] only include data up until the first quarter of 2020 due to the complex geopolitical and economic impact of the COVID-19 pandemic. This balanced panel dataset consists of 1820 observations, where each row is a bank. The *Treatment* variable is assigned a value of 1 to banks with *Total Consolidated Assets* between \$50 billion and \$250 billion and 0 otherwise. The *Post* variable takes a value of 0 if the observation of a particular bank is before quarter 1 of 2018 and takes a value of 1 if it is after quarter 1 of 2018 (passage of the EGRRCPA). The variables are all winsorized at the 1^{st} and 99^{th} percentile and standard errors clustered at the bank-level for the baseline

regressions. In total, 91 unique banks and 20 quarters are included in the analysis.

Chapter 4

Methodology

4.1 Difference-in-differences

Our baseline findings are based on a difference-in-differences approach with TWFE to control for variations within the banks over time. As mentioned above, we evaluate the treatment effect, β, on the outcome Y, which is the log-change of *Risk Weighted Assets* (RWA) for a particular bank, ΔRWA_{it}. D_{it} is the interaction term of banks in the treatment group and banks in the post-treatment period. X represents the series of controls (*Deposit Funding*, *Provisions Ratio*, *Operating Efficiency*, *Dividend Payout Ratio*, *Derivatives Ratio*, and *Liquidity Ratio*). γ_i is the bank-level fixed effects to capture unobserved heterogeneity, δ_t is the time-level fixed effects capturing common macroeconomic or regulatory shocks, and ϵ_{it} is the error term, clustered at bank level.

$$Y_{it} = \alpha' X_{it} + \beta D_{it} + \gamma_i + \delta_t + \epsilon_{it} \tag{4.1}$$

Under the assumption that all time-varying differences between the treatment and the control group are accounted for, the difference-in-differences is a popular and easily interpretable method used to identify causal relationships between the timing of an intervention and its effect on the treated group. This is done by comparing the outcomes before and after the policy went into effect between the treated and control groups to estimate the causal effect.

At the most basic level, it can be described as four averages and three differences. This is done by calculating the average observed outcome for the treated group before the treatment ($\bar{y}_{T0}$), the average observed outcome for the treated group after the treatment ($\bar{y}_{T1}$), the average observed outcome for the control group before the treatment ($\bar{y}_{C0}$), and the average observed outcome for the control group after the treatment ($\bar{y}_{C1}$). The "difference in differences" is then calculated by subtracting the difference in outcomes for the control group from the difference of the treated group: $(\bar{y}_{T1} - \bar{y}_{T0}) - (\bar{y}_{C1} - \bar{y}_{C0})$. Before adjusting for controls, the regression coefficient is equal to the manual difference-in-differences estimate obtained by subtracting these four means.

However, a key shortcoming of traditional causal methods is that they often require strong assumptions that may not hold true in practice. For example, the strict assumption of parallel trends in difference-in-differences is one such limitation. This assumption posits that, in the absence of the treatment, the treated and control groups would have followed the same trend over time. Violations of this assumption can lead to biased estimates of the treatment effect. In reality, the treated and control groups might experience different pre-treatment trends, which makes causal analysis more difficult.

Additionally, standard causal methods often struggle with handling a high-dimensional set of controls. When there are many covariates, issues such as multicollinearity can arise, making it difficult to accurately estimate the treatment effect. Furthermore, these models often assume a specific functional form for the relationships between variables (e.g., linearity), which can lead to misspecification and biased estimates if the true relationship is more complex. Traditional methods may also rely on strong assumptions about the data, such as exogeneity in instrumental variable models or the aforementioned parallel trends assumption, which can limit the reliability of the causal estimates.

To address these issues, we use a technique known as double machine learning, proposed

by Chernozhukov et al. [10], which is a causal technique that uses machine learning to allow for greater flexibility and robustness in the face of high-dimensional data, complex relationships among covariates, and potential violations of traditional assumptions. DML is a modern approach that leverages machine learning algorithms to flexibly estimate models for predicting the outcome from the covariates and for predicting the treatment assignment from the covariates. The estimates from these two models (the treatment assignment model and outcome model) are known as nuisance parameters. DML then uses these estimates to compute the treatment effects. By doing so, it provides doubly-robust estimates, meaning it remains consistent even if either of the nuisance models is slightly misspecified, as long as the other is correctly specified.

4.2 Double Machine Learning

Traditional machine learning aims to make accurate predictions by minimizing predictive error (e.g., mean squared error) on a test set. The focus is on improving the model's ability to generalize to unseen data, often through techniques like cross-validation and regularization. In this context, the goal is purely predictive, understanding the relationship between input variables (features) and the target outcome, with little to no concern for causality. In contrast, causal analysis is focused on understanding the relationships between the treatment, outcome, and covariates in order to estimate causal effects. Here, the primary objective is not prediction but to identify how changes in the treatment variable influence the outcome, while controlling for confounding factors.

If traditional machine learning methods are applied directly in causal contexts without adjustments, they may yield biased estimates of treatment effects. This is because such methods are typically designed to optimize for predictive accuracy, which may involve regularizing

aggressively, inducing bias, and therefore not capturing patterns in the data that reflect true causal relationships. In particular, machine learning models do not generally account for endogeneity or confounding between the treatment, outcome, and covariates, leading to biased causal inference.

The method of DML developed by Chernozhukov et al. [10] is designed to bridge the gap between causal inference, and the predictive strengths of machine learning, particularly in handling non-linear relationships and high-dimensional data. It does so by leveraging the concept of Neyman orthogonality, a mathematical condition that ensures that small errors in the nuisance estimation do not bias the moment conditions used to estimate the treatment effect. This is important as machine learning is sensitive to hyperparameter choice. Using moment conditions and cross fitting, Chernozhukov et al. [10] show that the causal estimate is root-N consistent, meaning that as the sample size increases the estimator becomes more accurate and its sampling distribution converges to a normal distribution.

DML is often cited as a method to reduce the bias that is introduced by traditional machine learning techniques. By implementing the process described later in this section, DML helps to generate unbiased estimates of the treatment effect even when the data itself is not high dimensional. Most of the current literature using DML is applied in the cross-sectional data setting. So far, little empirical work has been done to demonstrate the viability of using DML in panel data settings, and its extension to panel data remains non-trivial. Due to the nature of panel data tracking the same units over time, more complexities arise as compared to the cross-sectional setting.

The cross-fitting scheme, is used to ensure valid out-of-sample machine learning predictions. Typically, the data is split the data into K folds: the models are trained on $K-1$ folds and evaluated on the remaining k^{th} fold which is a test set. This cross-fitting procedure becomes more challenging in the panel setting due to the dependency of the same units across multiple

time periods. The random splitting used in the traditional DML implementation relies on the assumption that observations are independently and identically distributed. Additionally, accounting for the unobserved differences between the units and time periods, known as unobserved heterogeneity, becomes difficult in the panel setting. Unobserved heterogeneity refers to the differences between banks and across time that we cannot measure directly, but still affect the outcome. For example, different banks may have different organizational cultures or management styles that are immeasurable. Similarly, over the course of several years, there may be broader macroeconomic conditions that the control variables are unable to account for. Addressing the concept of unobserved heterogeneity, is crucial for the adaptation of DML to panel data. In traditional linear regression models, as shown in (4.1), we account for these immeasurable characteristics through fixed effects typically represented using dummy variables for bank fixed effects and time fixed effects. Using machine learning also aims to significantly help capture complex non-linear relationships. In the panel setting, fixed effects are used to capture the time-invariant heterogeneity, but incorporating these fixed effects for machine learning models is a challenge. Furthermore, time-varying heterogeneity is harder to handle because it may still be correlated with both the treatment and outcome variables.

In the rest of this section we discuss the two main estimation settings of DML, the partially linear and the interactive models. We then discuss the data transformations needed to address the unobserved heterogeneity and potential choices for cross-fitting. In the following section, we describe the recent approach by Clarke and Polselli [13] which modifies the partially linear setting for panel data. Finally, we discuss machine learning estimators and hyperparameter choice for nuisance parameters.

4.2.1 Partially Linear Model

At its core, the partially linear setting DML consists of two nuisance parameters, the outcome and treatment, which are estimated using machine learning and have their residuals regressed to obtain the estimated treatment effect. In the partially linear model (4.2), where the treatment is conditionally exogenous, the treatment effect, θ, can be estimated the following way. The data is split into K random folds to predict the nuisance parameters, $m(X)$ and $\ell(X)$, using machine learning, where $m(X) = \mathbb{E}[Y|X]$ and $\ell(X) = \mathbb{E}[D|X]$. The procedure, known as cross-fitting, involves using $K-1$ folds as the training set and predicting the nuisance parameters on the remaining held-out fold. These estimates are appended to their respective prediction vectors. After the estimation of the outcome and treatment nuisance parameters, we calculate their residuals by taking the difference of the prediction vectors and the true values (4.3). We then use these residuals to estimate the treatment effect, θ, by regressing the residuals of the treatment predictions on the residuals of outcome predictions (4.4). The random cross-fitting procedure, similar to K-fold cross validation, splits the data into folds to ensure that the nuisance parameters are estimated out-of-sample to reduce bias and the residuals of the nuisance parameters ensure orthogonality with respect to the covariates, X.

$$Y = \theta D + \ell(X) + U, \tag{4.2}$$

$$\begin{aligned} \tilde{Y} &= Y - \hat{m}(X) \\ \tilde{D} &= D - \hat{\ell}(X), \end{aligned} \tag{4.3}$$

$$\tilde{Y} = \theta \tilde{D} + \epsilon. \tag{4.4}$$

This is represented more formally, through the Neyman orthogonal moment condition expressed as the "partialling-out" score function by Chernozhukov et al. [10]. This is a formalized way to describe the process of learning and estimating the nuisance parameters $\ell(X)$ and $m(X)$.

$$\psi(W;\theta,\eta) := \{Y - \ell(X) - \theta(D - m(X))\}(D - m(X)), \eta = (\ell, m), \tag{4.5}$$

4.2.2 General Interactive Model

Due to the nature of our data and the task of analyzing policy on a set of banks, we also use a more general model, known as the interactive model, to flexibly account for the heterogeneity in the treatment. Since the treatment corresponds to the implementation of a policy, a binary intervention that becomes effective after a specific time period for banks of a certain size, its effect may vary across banks and over time. For instance, larger banks might respond differently to policy changes than smaller ones. Furthermore, the relationship between bank size, policy implementation, and outcomes might not be strictly linear. As a result, the interactive model offers greater flexibility in modeling this relationship. In contrast, the partially linear model assumes constant treatment effects and a simple additive relationship between the treatment, covariates, and outcome, which may be less nuanced due to the potential variations in how different banks respond to the policy. In this context, the interactive model allows for the possibility that the impact of the policy depends on both individual bank characteristics and time-varying factors. We adopt the general interactive model as introduced by Chernozhukov et al. [10], which enables us to account for treatment effect heterogeneity. This model is a flexible variation of the partially linear model in equation (4.2). Unlike the partially linear model, which requires that D be exogenous, the general

interactive model does not rely on random treatment assignment. Instead, it is structured such that the nuisance parameter $g(X)$ captures both the treatment and its interaction with the covariates, acknowledging that D is not additively separable.

$$Y = g(D, X) + U, \quad \mathbb{E}[U|X, D] = 0 \tag{4.6}$$

$$D = m(X) + V, \quad \mathbb{E}[V|X] = 0. \tag{4.7}$$

Chernozhukov et al. [10] establishes the following score functions to calculate the average treatment effect (ATE) and average treatment effect on the treated (ATTE), based on the potential outcomes (4.8) and (4.9) respectively. We use the residuals and the nuisance functions within the context of these score functions to estimate the ATE (4.10) and ATTE (4.11), where $\mathbb{E}_n[D]$ is the treatment probability, $\mathbb{E}[g(1, X)]$ is the estimated outcome for the treated, and $\mathbb{E}[g(0, X)]$ is the estimated outcome for the untreated. The terms involving $m(X)$ act as weights to correct for the fact that we do not observe both potential outcomes for each individual. To solve, set $\mathbb{E}[\psi(W; \theta, \eta)] = 0$. We show how to empirically solve for the causal parameter θ in Appendix B.

$$\theta = \mathbb{E}[g(1, X) - g(0, X)], \tag{4.8}$$

$$\theta = \mathbb{E}[g(1, X) - g(0, X)|D = 1], \tag{4.9}$$

$$\psi(W;\theta,\eta) := (g(1,X) - g(0,X)) + \frac{D(Y - g(1,X))}{m(X)} - \frac{(1-D)(Y - g(0,X))}{1 - m(X)} - \theta, \quad (4.10)$$

$$\psi(W;\theta,\eta) := \frac{D(Y - g(0,X))}{\mathbb{E}_n[D]} - \frac{m(X)(1-D)(Y - g(0,X))}{\mathbb{E}_n[D] \cdot (1 - m(X))} - \frac{D\theta}{\mathbb{E}_n[D]}, \quad (4.11)$$

4.2.3 Panel Data Transformations

As stated earlier, one of the primary concerns when dealing with panel data and applying it to this causal machine learning framework is accounting for the unobserved heterogeneity. To account for the unit-level heterogeneity and temporal dependence faced in the panel setting, we use dummy variables, demean, first-difference, and use correlated random effects (CRE).

Demeaning and first-differencing both transform the data. Demeaning allows us to control for both unit and time fixed effects. Suppose W_{it} is a vector of observed variables for unit i at time t, which could represent the outcome Y_{it}, treatment D_{it}, or covariates X_{it}. We define the unit-specific mean $\bar{W}_i = \frac{1}{T}\sum_{t=1}^{T} W_{it}$, the time-specific mean $\bar{W}_t = \frac{1}{N}\sum_{i=1}^{N} W_{it}$, and the overall grand mean $\bar{W} = \frac{1}{NT}\sum_{i=1}^{N}\sum_{t=1}^{T} W_{it}$. The two-way demeaned variables are constructed such that $\tilde{W}_{it} = W_{it} - \bar{W}_i - \bar{W}_t + \bar{W}$.

First-differencing aims to eliminate the unobserved unit-specific effect by taking the difference of adjacent time periods. However, because no lagged observation exists for the first time period, the differenced dataset only contains observations from $t = 2, \ldots, T$. Using W_{it}, it is defined as $\Delta W_{it} = W_{it} - W_{it-1}$.

Correlated random effects and dummy variables are approaches that expand the set of covariates. CRE expands the set of controls so that the estimation and training of the model

include the original raw covariates and their unit and/or time-specific means. We evaluate the CRE approach twice: first, using both unit and time-specific means in the covariate set $\{X_{it}, \bar{X}_i, \bar{X}_t\}$, and then using just the unit-specific means $\{X_{it}, \bar{X}_i\}$. Going forward we mention both these methods as CRE and CRE-unit respectively. Dummy variables also augment the set of covariates by adding a binary variable for each unit and time period that takes the value of 1 for the rows that fall in the unit and/or time category denoted by that dummy column, and 0 otherwise.

Cross-fitting for Panel Data

The concept of a training and testing set is crucial in machine learning to reduce bias by ensuring that the models generalize well enough to account for new unseen data. The training set is the input that the machine learning models are initially trained on and the test or validation set is the subset of the data that the model uses to make predictions. As stated earlier, one of the primary challenges when dealing with panel data in the context of a causal machine learning framework is designing a cross-fitting procedure. If we split the data randomly (as is the case in traditional DML), these temporal and unit-level relationships are not captured, meaning that the observations are not truly independent. To overcome this challenge we can split the data in a couple of ways. We can split the data such that all the observations of the same unit are contained in the same fold. However, in this case, the machine learning models will be unable to predict any unit specific effects. Similarly, when the data is split such that all the observations within the same time period lie within the same fold, models will be unable to predict any time specific effects on the out-of-sample folds. Random splits of the data could be used to break these dependencies, but could lead to biased estimates. Fuhr and Papies [17] and Clarke and Polselli [13] explicitly address modifying the cross-fitting procedure for the panel data setting. We apply two main splitting schemes.

The first is a random split, as originally proposed in Chernozhukov et al. [10]. Secondly, we use the technique laid out by Clarke and Polselli [13] which splits the data by unit, ensuring that all observations for a given unit are allocated to the same training fold. This is called the *block-k-fold* cross-fitting procedure. It's interesting to note that Fuhr and Papies [17], through multiple data generating processes and Monte Carlo simulations, find the choice of splitting procedures to have little impact on the causal parameter. Each of the latter two methods addresses dependency along one dimension but leaves the other unaccounted for. For instance, when splitting by unit, temporal dependence between observations is not accounted for in the cross-fitting procedure. Similarly, when splitting by time period, dependence between observations within the same unit remains unaddressed. Consequently, residual dependencies may still persist in the splits, which could impact the validity of the causal inference.

The blocked cross-fitting incurs more of an overhead cost due to the need to first identify the unique groups and then split the data into each fold by group rather than by individual observation. Interestingly, from an algorithm analysis perspective, the efficiency of both the random cross-fitting procedure and the block-k-fold cross-fitting procedure have an upper bound of $O(n)$. This is due to the fact that, as the input size increases, the operations scale linearly. Whether the cross-fitting is being done randomly at the observation level or block-k at the unit level, every row is evaluated only once to assign it to a fold. The training time and prediction times remain the same, with the only difference being the overhead of keeping all unit observations in the same fold for the block-k cross-fitting. Even though this overhead increases the constants and makes the blocked cross-fitting more expensive in practice – especially as the number of units grows large – it does not increase the asymptotic rate of growth of the algorithm, as it still scales linearly under both random and block-k schemes. However, in block-k cross-fitting, practical costs grow noticeably as the number

of units increases, due to the need to distribute large or uneven units across folds and the added expense of indexing and grouping operations. These practical costs can impact wall-clock times and resource usage when considering the full DML pipeline, especially in high-dimensional settings when the panel data is imbalanced. Figure 4.1 shows the execution time for generating 5 folds through the random and block-k-fold schemes. As expected, they both seem to grow linearly as the number of units grows large. Additionally, we see that the block-k method (labeled as *Grouped* since we use the GroupKFold method from scikit-learn for this test) takes slightly longer due to the additional overhead.

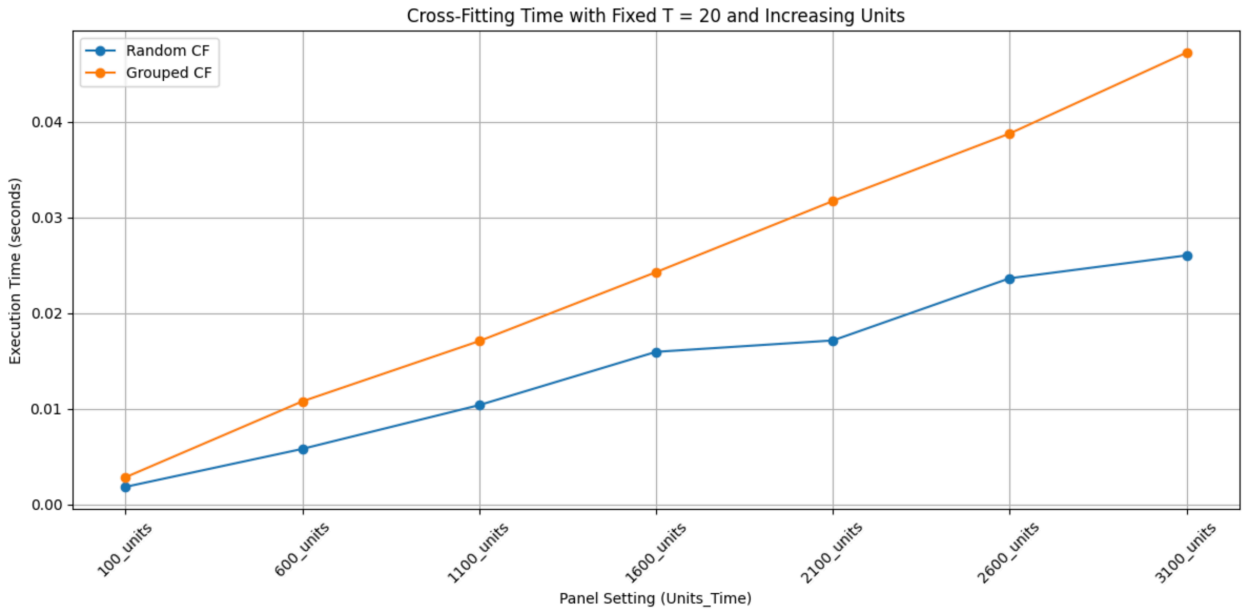

Figure 4.1: Execution time of generating 5 folds during random and grouped cross-fitting procedures on a balanced panel dataset as the number of units grows large. We only time the fold generation step. This does not include training any machine learning models or estimating the nuisance functions.

4.3 Partially Linear DML for Panel Data

Clarke and Polselli [13] recently proposed a new method for the estimation of causal effects

in panel data. They introduce the concept of block-k-fold cross fitting, which involves sample splitting the data such that all observations for a unit are fully contained within a single fold. As noted earlier, this helps to satisfy the independent and identically distributed assumption, but has the drawback of not being able to predict unit-specific effects in the held out fold. Additionally, they demonstrate that the use of CREs and transforming the data through first-differencing and demeaning are viable techniques for addressing the unit and time fixed effects. Clarke and Polselli [13] implement their estimating procedures in R through the XTDML library that they built on top of the DoubleML library from Bach et al. [2]. For the estimation of the treatment effects, they adjust the Neyman orthogonal score functions defined by Chernozhukov et al. [10] for the partially linear setting. The steps to estimating the target causal parameter, θ, are almost identical to the steps laid out by Chernozhukov et al. [10] except with modifications to account for the panel data. This involves performing the block-k-fold cross-fitting by randomly partitioning the cross-fitting units into K folds. This is followed by learning the nuisance functions for the treatment and outcome on an individual fold and solving the Neyman orthogonal moment condition to obtain the estimate for that fold. These last two steps of learning the nuisance functions and estimating the causal parameter are repeated and averaged across all K folds.

The assumptions of this method closely follow those of the original partially linear DML procedure. Specifically:

1. **Covariate Independence:** Conditional on unit fixed effects and past covariates, current covariates are independent of past outcomes and treatments.

2. **Sequential Independence:** Conditional on current covariates and unit fixed effects, current outcomes and treatments are independent of past covariates, outcomes, and treatments.

3. **Unconfoundedness:** Conditional on observed covariates and fixed effects, treatment assignment is independent of potential outcomes (i.e., treatment is as good as randomly assigned).

4. **Constant Treatment Effect:** The treatment effect is assumed to be constant across time and units, meaning it does not vary with covariates or fixed effects. Note that this specifically applies for the partially linear model as the interactive model relaxes this assumption.

Below, we explain the four different approaches proposed by Clarke and Polselli [13] for the training and estimation of machine learning models for the nuisance functions of the treatment and outcome on panel data.

4.3.1 Correlated Random Effects Approach

As the name implies, the correlated random effects approach establishes a procedure for learning the nuisance functions of the treatment and outcome by using CREs to account for fixed effects. Suppose that t represents time periods and i represents units, which in our case is the number of quarters the data was collected (Q2 2015 to Q1 2020) and the number of banks respectively. For illustrative purposes, consider the case in which we examine the CRE-unit (only unit-level means). Of the two nuisance functions that need to be estimated to obtain the causal parameter, the nuisance function for the outcome, $\tilde{\ell}(X_{it}, \bar{X}_i)$, is trained on the data $\{Y_{it}, X_{it}, \bar{X}_i\}$. The authors go on to explain that it is incorrect to simply train the nuisance function for the treatment, $\tilde{m}(X_{it}, \bar{X}_i)$, on the data $\{D_{it}, X_{it}, \bar{X}_i\}$, and then obtain the residuals from the treatment vector like the typical partially linear model: $\hat{V}_{it} = D_{it} - \hat{m}(X_{it}, \bar{X}_i)$. This partly ignores the unit-level random effect present in the treatment mean, c_i. This is important as c_i represents the individual unit-level characteristics

that affect the outcome but are not fully included in $\{X_{it}, \bar{X}_i\}$. The reason that the treatment nuisance function cannot be trained on the data $\{D_{it}, X_{it}, \bar{X}_i, \bar{D}_i\}$ is due to the fact that unit-level treatment mean, $\bar{D}_i$, naturally contains information about the treatment, D_{it}. Including $\bar{D}_i$ as input in the training phase would likely result in data leakage and biased nuisance functions. Additionally, the partially linear model does not allow for the treatment to be included as input for the models.

To rectify these issues, the approach proposed by Clarke and Polselli [13] involves using the unit-level treatment mean, $\bar{D}_i$, in the prediction phase after the model has already been trained on the data $\{D_{it}, X_{it}, \bar{X}_i\}$. More precisely, this approach is broken down into three steps. First, the machine learning model, $\tilde{m}(X_{it}, \bar{X}_i)$, is trained on the data $\{D_{it}, X_{it}, \bar{X}_i\}$. This uses only covariates and their unit-level means as input. Second, the average predicted treatment for each unit is calculated as $\bar{m}_i = \frac{1}{T}\sum_{t=1}^{T} \tilde{m}(X_{it}, \bar{X}_i)$. The final corrected model used in the prediction phase is defined as $m^*(X_{it}, \bar{X}_i, \bar{D}_i) = \tilde{m}_i(X_{it}, \bar{X}_i) + (\bar{D}_i - \bar{m}_i)$, where the first term is just the original model trained only on $\{D_{it}, X_{it}, \bar{X}_i\}$ and the second term is the correction term that takes the difference between the unit-level treatment mean and the average predicted treatment for each unit. This second term essentially acts as a proxy for the unobserved term, c_i.

4.3.2 Approximate Approach

The approximate approach involves the use of transforming the data through either demeaning or first-differencing before the training and estimation phase. If we suppose that Q represents the transformation of the data then nuisance functions for the treatment and outcome can be learned from the data $\{Q(D_{it}), Q(X_{it})\}$ and $\{Q(Y_{it}), Q(X_{it})\}$. Clarke and Polselli [13] state that this approach is not as robust and generally inferior to the other approaches as

it only works in a linear setting. More specifically, it assumes that $Q(m(X_{it})) = m(Q(X_{it}))$ and $Q(\ell(X_{it})) = \ell(Q(X_{it}))$, which only holds when ℓ and m are linear.

4.3.3 Exact Approach

This approach is the most robust as it allows for the estimation of the nuisance functions to represent the conditional expectation of the transformed variables. This technique involves the covariates and their lags $\{X_{it}, X_{i,t-1}\}$ as opposed to the first-differenced data $\{\Delta X_{it}\}$. The model for the treatment and outcome and trained and estimated using the data $\{\Delta D_{it}, X_{it}, X_{i,t-1}\}$ and $\{\Delta Y_{it}, X_{it}, X_{i,t-1}\}$ respectively. This ensures that the learned model corresponds to the transformed data in non-linear settings.

4.3.4 Hybrid Approach

This is a hybrid procedure that involves learning and generating estimates for the nuisance functions using CREs with the raw covariates, $\ell(X_{it}, \bar{X}_i)$ and $m(X_{it}, \bar{X}_i)$. This is followed by a post-estimation transformation that transforms the predictions, through either first-differencing or time demeaning, $Q(\ell(X_{it}, \bar{X}_i))$ and $Q(m(X_{it}, \bar{X}_i))$, where Q represents the transformation of the data.

4.4 Nuisance Estimators

We use two tree-based and two linear machine learning methods. These include extreme gradient boost, random forest, a combination of ordinary least squares (OLS) and logistic regression, and lasso. We employ a 5-fold cross-fitting scheme for the DML nuisance estimations and, where applicable, use random search to optimize hyperparameters for the random

forest and gradient boosted learners.

4.4.1 Hyperparameter Choice

Several considerations had to be made regarding how to tune hyperparameters of the machine learning models for the nuisance parameters. Due to the sample splitting nature of DML, where the data is split into random folds and then the nuisance parameters are estimated on the test set of each fold, we could tune the model on each fold. However, this method is computationally expensive and time consuming. To find an alternative, we follow the guidance of Bach et al. [3], who show that tuning on the full dataset produces comparable results. We specify model parameters for the gradient boost and random forest models and evaluate them based on a random search through a 5-fold cross validation procedure. For the interactive and partially linear setting, we use the XGBoost and scikit-learn libraries in python to implement the extreme gradient boost and random forest models. We configure a classifier to estimate the treatment nuisance parameter and a regressor for the outcome nuisance parameter for each algorithm. This results in the treatment nuisance function being modeled with a random forest classifier and XGBoost classifier, and the outcome nuisance function being modeled with a random forest regressor and XGBoost regressor. Since the demeaning transformation results in a non-binary treatment vector, we use the XGBoost regressor and random forest regressor for all machine learning models on the demeaned data.

The hyperparameters for both models are optimized using 100 iterations of random search to ensure a comprehensively tuned and validated model. The search space for the random forest classifier in python includes: *n_estimators* (randomly drawn from 500 to 1000), *max_depth* (randomly drawn from 2 to 20), and *min_samples_leaf* (randomly drawn from 1 to 10).

The random forest regressor employs an identical search space. We used R to implement the CRE, exact, and approximate approach as proposed by Clarke and Polselli [13]. The same search space is used in R: *num.trees*, *max.depth*, and *min.node.size*. The random forest regressor employs an identical search space.

For the XGBoost algorithm, we utilized a similar approach to hyperparameter tuning for both the classifier and regressor. The search space in python included: *n_estimators* (randomly drawn from 500 to 1000), *learning_rate* (randomly drawn from 0.01 to 0.2), *max_depth* (randomly drawn from 2 to 10), and *gamma* (randomly drawn from 0 to 0.5). The same search space is used in R: *nrounds*, *eta*, *max_depth*, and *gamma*. The XGBoost regressor employs an identical search space.

To incorporate regularized linear regression into our analysis, we utilized both lasso and ridge regression. In Python, we employed the scikit-learn implementations of LassoCV and RidgeCV, while in R, we leveraged the glmnet package. We use cross-validation to determine the optimal regularization parameter, lambda.

In addition to the machine learning we also employ linear models for the nuisance function estimation. Again, since the demeaned data results in non-binary treatments, we use OLS instead of the logistic regression for the treatment nuisance function of the demeaned data. For all other transformations and approaches (dummy, CRE, and first-differencing) we use logistic regression as the model to estimate the treatment nuisance function and OLS to estimate the outcome nuisance function.

Although we perform random cross-fitting with 100 iterations, the mathematical condition of Neyman orthogonality and the cross-fitting procedure help to ensure that small estimation errors in nuisance functions do not bias the estimation of the causal parameter. The condition of Neyman orthogonality is important, as machine learning models are sensitive

to hyperparameter choice. Chernozhukov et al. [10] rigorously prove that the orthogonal score functions yield moment conditions that are locally insensitive to small errors in the nuisance parameters. Due to this, we believe it is sometimes preferable to allow slightly more overfitted machine learning models for the nuisance estimations, rather than overly regularizing.

Chapter 5

Results

5.1 Baseline Findings

We run a classic TWFE model with fixed effects for bank and time, and standard errors clustered on the bank level. We run this on the same data from Chronopoulos et al. [12]. We display the results of linear regressions using dummy variables (traditional fixed effects), CRE-unit, and CRE with unit and time means. We also show the results of the two transformation approaches: demean and first-difference. In this linear setting we obtain significant treatment effects for all approaches. We obtain similar results to the TWFE effects model from Chronopoulos et al. [12]. We use a balanced panel dataset with all available quarters from quarter 2 2015 to quarter 1 2020. We find *Operating Efficiency*, *Dividend Payout Ratio*, and *Liquidity Ratio* to be the most significant controls.

Parallel Trends

An important assumption under this setting is that of parallel trends. We can see a weak but visible parallel trend. In the plot below we run a fixed effects regression with the treated group and time dummy interactions. In 5.1, if the pre-treatment coefficients are close to zero and not statistically significant, it supports the idea that the treated and control observations were on similar trends before the intervention.

Table 5.1: Baseline Regression Results

	Dummy	Demean	CRE Unit	CRE	First Difference
Post × Treated	0.0193 (0.0045) ***	0.0193 (0.0044) ***	0.0184 (0.0037) ***	0.0194 (0.0044) ***	0.0189 (0.0054) ***
Deposit Funding	-0.0105 (0.0326)	-0.0105 (0.0316)	-0.0176 (0.0302)	-0.0174 (0.0305)	-0.0934 (0.1079)
Provisions Ratio	0.0110 (0.0175)	0.0110 (0.0169)	0.0015 (0.0128)	0.0111 (0.0168)	0.0282 (0.0184)
Operating Efficiency	0.2383 (0.0514) ***	0.2383 (0.0498) ***	0.2397 (0.0455) ***	0.2358 (0.0494) ***	0.3441 (0.0615) ***
Dividend Payout Ratio	-0.0529 (0.0294) *	-0.0529 (0.0285) *	-0.0547 (0.0302) *	-0.0542 (0.0299) *	-0.0621 (0.0389)
Derivatives Ratio	0.0176 (0.0137)	0.0176 (0.0132)	0.0156 (0.0131)	0.0171 (0.0131)	-0.0166 (0.0314)
Liquidity Ratio	-0.1281 (0.0661) *	-0.1281 (0.0641) **	-0.1453 (0.0645) **	-0.1314 (0.0642) **	-0.2953 (0.1009)***
Observations	1820	1820	1820	1820	1729
Adj. R2	0.1610	0.0959	0.0946	0.1102	0.1361
Total Covariates	117	7	14	21	7

Significance levels indicated by '***', '**', '*' for the 1, 5, and 10 percent levels respectively. Standard errors are clustered at the bank level. Note that the *Total Covariates* includes the treatment and any other variables included as a predictor.

5.2 DML - Panel Setting

Few recent applied works in DML aim to incorporate the method in a panel data framework. Yuan and Liu [31] and Gao et al. [18] measure the impact of urbanization on Chinease cities using DML with panel data. However, they implement the three-stage procedure with the final stage as an instrumental variable regression. Additionally, they do not provide any detail regarding the data transformations that they implement to adjust for the panel data. We include five different approaches for handling the data, as described earlier. We demean the data by unit and time, incorporate correlated random effects with unit means, correlated random effects with unit and time means, and first-differencing. We also run DML on data that has not been adjusted in any way whatsoever. This is denoted as *Pooled*, following the notation of a pooled OLS, which is a panel data regression where fixed effects are unaccounted

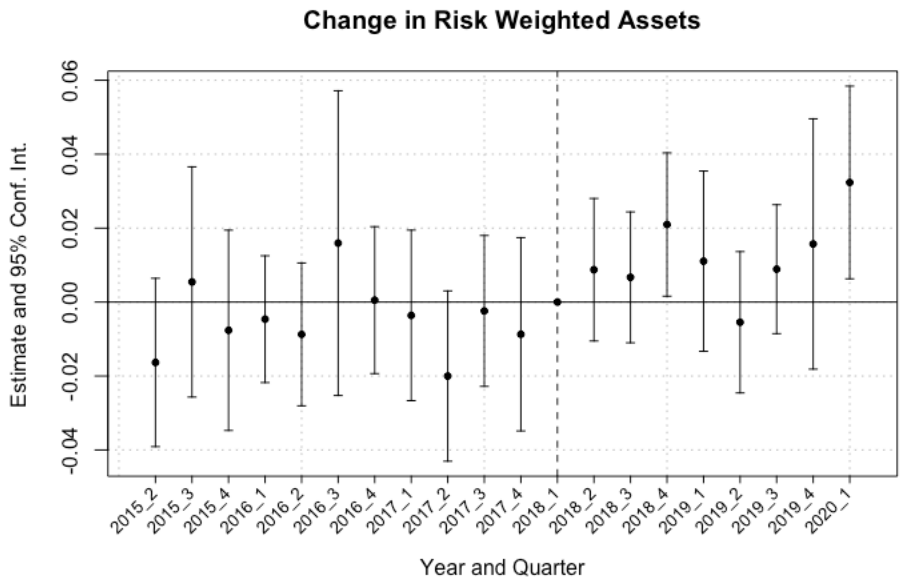

Figure 5.1: Parallel trends plot showing the dynamic treatment effects by interacting time dummies with the treatment indicator.

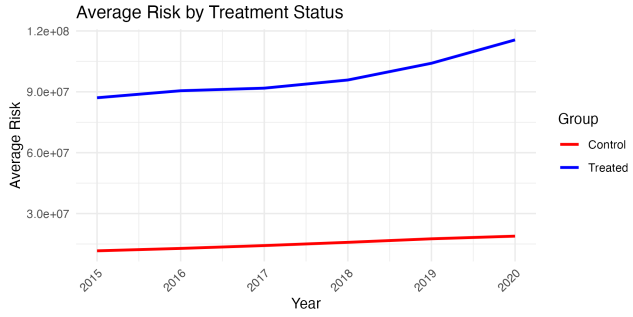

Figure 5.2: The average *Risk Weighted Assets* per year across banks in the treatment and control group.

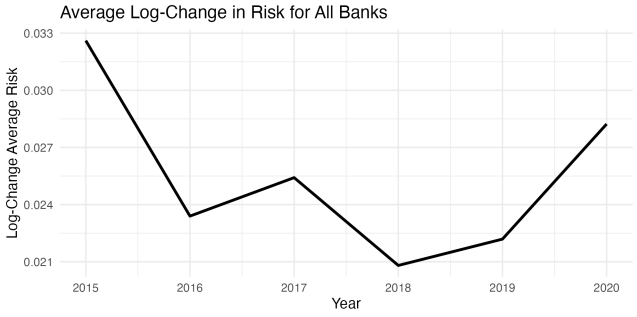

Figure 5.3: Combined log-change in *Risk Weighted Assets* averaged across all banks over the years.

for. For the partially linear setting, we obtain treatment effect estimates from the DoubleML library which uses the Neyman orthogonal score function. This is effectively the ATE for the partially linear model. We include results of the interactive model for both the ATE and the ATTE. We investigate the interactive model to account for the treatment effect heterogeneity and relax functional assumptions on our covariates. This treatment effect is estimated using the score functions mentioned earlier, (4.10) for the ATE and (4.11) for ATTE.

We first implement a naive DML approach to estimate the causal parameters under the partially linear model and interactive model settings using the standard DoubleML library. In this setup, we directly apply machine learning estimators while only partially accounting for the panel structure of the data. Specifically, we use the default random cross-fitting and naively estimate the transformed data, similar to the approximate approach by Clarke and Polselli [13]. We first show the results of the partially linear and interactive models with just the raw data, not accounting for any panel structure (*Pooled*). Even after we include transformations, the default practice of random cross-fitting seems to yield poor results. Thus, while we address the panel data through transformations and CRE, the cross-fitting design leads to leakage and does not respect the dependency across units over time. Even after applying the transformations, the naive DML does not fully implement the corrected nuisance function structure required for panel data. For instance, the naive method does not allow for the correction term in the CRE approach that uses the unit-level treatment mean as an input feature when predicting the treatment assignment (i.e., in the estimation of $m(X)$). This poses a threat of the treatment model being misspecified, particularly in the presence of treatment effect heterogeneity across units. Finally, the standard errors produced by DoubleML are not clustered at the unit-level, which poses additional concerns. If residual dependence exists within units over time, failing to account for it may result in biased standard errors and invalid confidence intervals. To address these limitations, we

subsequently extend the DML framework by incorporating methods from Clarke and Polselli [13] in the adjusted partially linear model.

5.2.1 Partially Linear Model

Unsurprisingly, we find the pooled estimates to be insignificant and poor, as the treatment effect is insignificant for most of the machine learning approaches. This was expected as there are no fixed effects. These confidence intervals are also biased because the naive method does not cluster standard errors. Additionally, we notice that the linear methods, lasso and OLS, have tighter confidence intervals. We find the demeaned results to provide the most significant and robust estimates for the machine learning methods, followed by the first-difference approach.

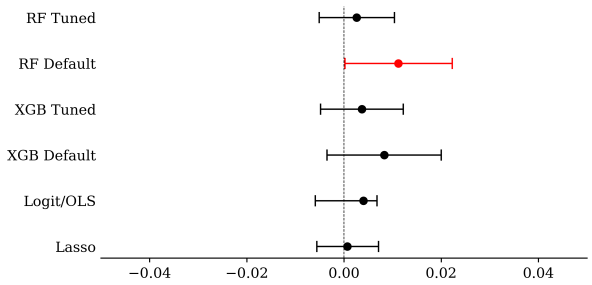

Figure 5.4: Naive PLM - Pooled

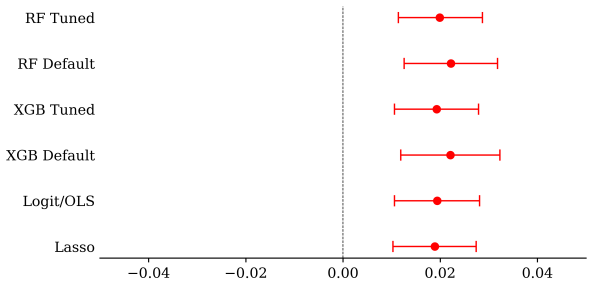

Figure 5.5: Naive PLM - Demeaned

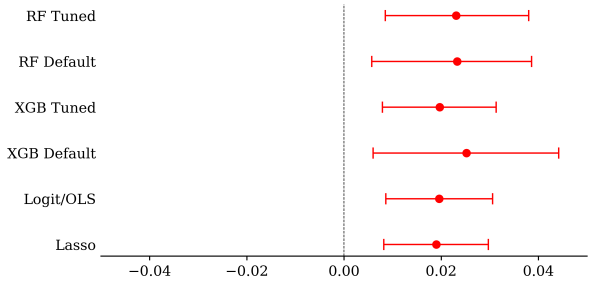

Figure 5.6: Naive PLM - First-difference

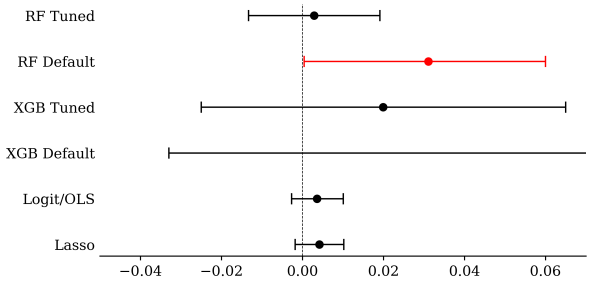

Figure 5.7: Naive PLM - CRE

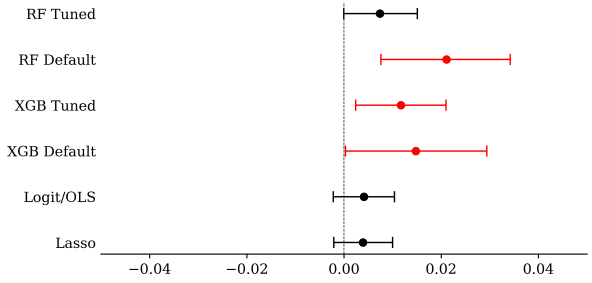

Figure 5.8: Naive PLM - CRE-unit

Even though the first-difference method has significant estimates, its confidence intervals are also more broad, indicating that they are not as stable as the demeaned or original linear models. We find that the machine learning techniques for CRE-unit display statistically significant estimates close to the original TWFE baseline of 0.193. Since we do not know the true causal estimate in practice, we follow the existing literature by evaluating the quality of the estimated coefficients by their magnitude among different learners and compare them using confidence intervals relative to other estimates. Fuhr and Papies [17] warn of strictly interpreting confidence intervals due to the existing heterogeneity and lack of full independence. Additionally, we try to maintain a consistent scale on the x-axis to allow for easier visual interpretability and comparison. We display the full estimates from the naive implementation in table A.1 in the appendix. In the estimate plots presented, we denote estimates that are significant in red. The plots are captioned as PLM and IM denoting the estimates obtained via the partially linear and interactive method, respectively.

5.2.2 Interactive Model

As stated previously, the interactive model allows for more expressive estimation of the outcome nuisance function by allowing the treatment to directly enter the feature set. For the naive interactive model, we begin by generating ATE and ATTE estimates from the DoubleML library.

We find that all approaches across machine learning methods tend to perform poorly in this setting. The estimates seem to vary wildly and are quite different from the estimate of 0.0193 obtained via TWFE. The ATE from the first-differencing procedure provides the most stable and significant estimates due to their similar estimates across both tuned and default machine learners, however, this result is also poor. Similar, to the naive partially

linear model, these estimates vary a lot and could be partly attributable to the lack of panel data preparation. Lastly, we note that the demean approach is infeasible for the interactive method, as Chernozhukov et al. [10] construct the interactive model for binary treatments only, and demeaning would invalidate this.

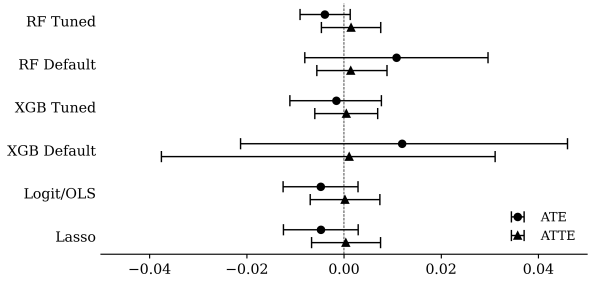

Figure 5.9: Naive IM - Pooled

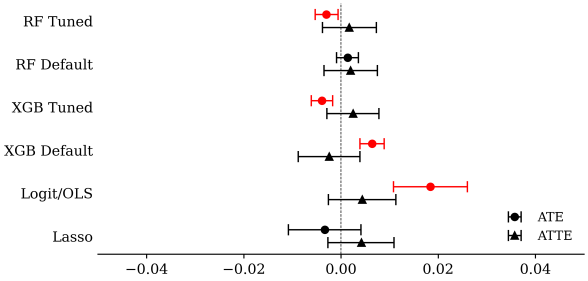

Figure 5.10: Naive IM - CRE

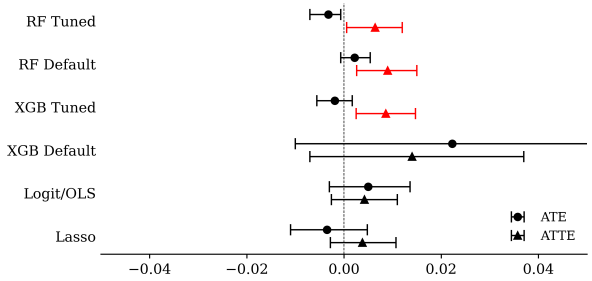

Figure 5.11: Naive IM - CRE-unit

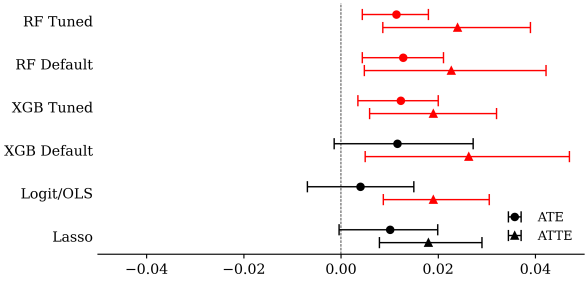

Figure 5.12: Naive IM - First-difference

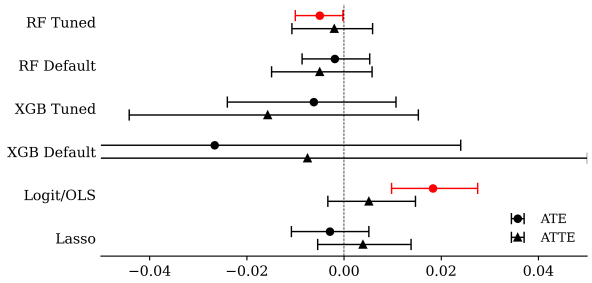

Figure 5.13: Clustered IM - CRE

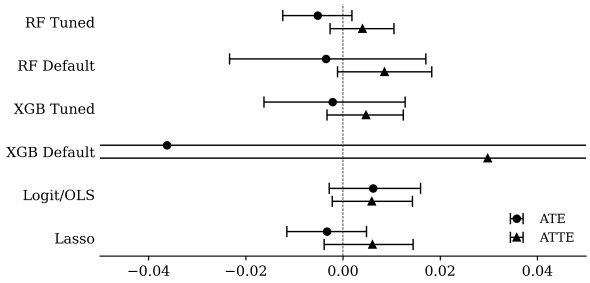

Figure 5.14: Clustered IM - CRE-unit

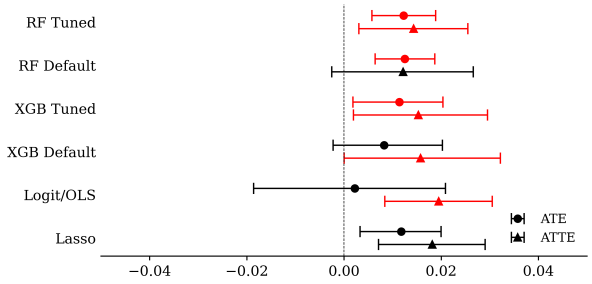

Figure 5.15: Clustered IM - First difference

We also perform block-k cross-fitting at the bank-level and plot the resulting estimates, denoted as *Clustered IM*. In the first-difference procedure we use the covariate lags, like Clarke and Polselli [13], instead of the actual differenced variables. However, these estimates are also quite inconsistent, as we do not notice much improvement from the naive interactive model to the clustered interactive model.

We display the full estimates from the naive and clustered implementation of the interactive model in table A.3 and A.4 in the appendix, respectively.

5.2.3 Adjusted Partially Linear Model

The adjusted partially linear model is the novel technique introduced by Clarke and Polselli [13]. We use the exact approach with first-differenced data, the approximate approach with demeaned data, and the CRE approach with unit means, and unit and time means. Ad-

ditionally, we expand the feature set through polynomial expansion to include interaction effects and potential nonlinear relationships among the covariates when using lasso. The polynomial expansion augments the original covariate matrix by generating additional features through second and third-order polynomial transformations. This is especially useful with penalized regression techniques like lasso, as the variable selection helps to manage the large number of covariates. Lastly, for the CREs, we use a regressor instead of a classifier for the treatment estimations with OLS (i.e., we do not use logistic regression) and lasso, as this yielded slightly more consistent results.

These results are theoretically more robust than the naive approaches above, as they fully account for the panel structure. Even with the adjusted model, we find that the CREs produce biased and unreliable estimates, regardless of the machine learning model employed. We find the first-differenced and demeaned data to provide the most consistent and significant estimates. The original TWFE estimate is shown in blue for comparison. The full table of estimates from the adjusted partially linear implementation is in table A.2 in the appendix.

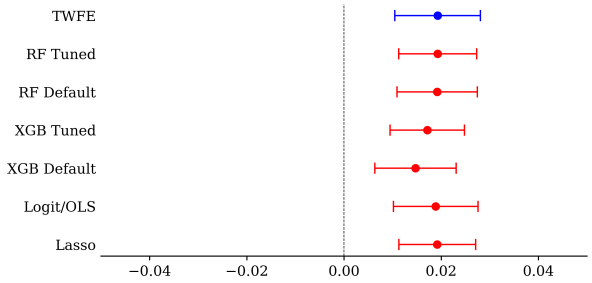

Figure 5.16: Adjusted PLM - Demeaned

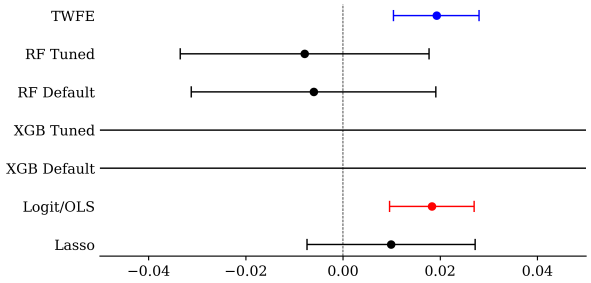

Figure 5.17: Adjusted PLM - CRE

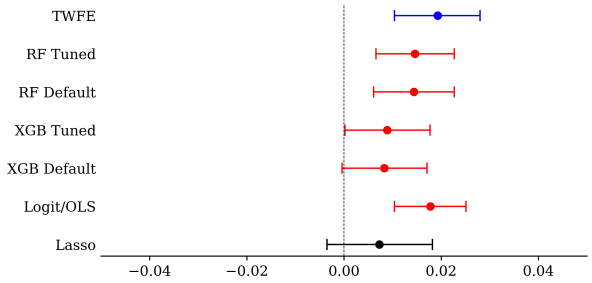

Figure 5.18: Adjusted PLM - CRE-unit

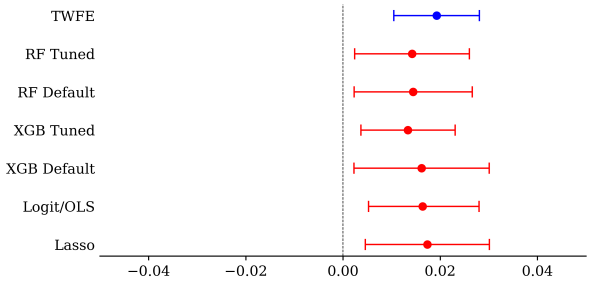

Figure 5.19: Adjusted PLM - First difference

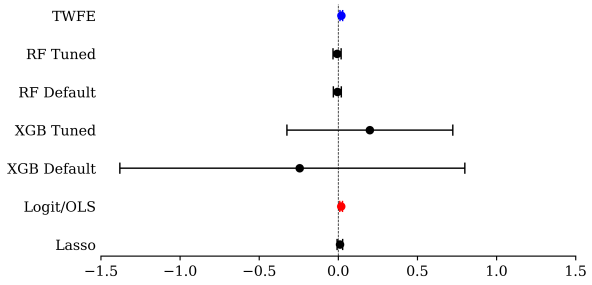

Figure 5.20: Adjusted PLM - CRE (widened x-axis)

Chapter 6

Discussion

Although the naive partially linear approach is consistent and robust under certain conditions, it does not account for panel structure in several key ways. Firstly, Clarke and Polselli [13] emphasize the importance of the exact method. In the first-differencing transformation, this is done by having the original covariates and their lags in the feature set for estimation and prediction of the nuisance functions. In contrast, the inputs we provide the DoubleML partially linear model for first-differencing are the fully transformed covariates such that the data used is $\{Y_{it} - Y_{it-1}, X_{it} - X_{it-1}\}$ for the outcome nuisance function and $\{D_{it} - D_{it-1}, X_{it} - X_{it-1}\}$ for the treatment nuisance function. This method also uses random cross-fitting, which can lead to leakage of unobserved heterogeneity across folds, potentially biasing the residuals and treatment effect estimates. Similarly, both the naive and clustered interactive method perform poorly and do not provide useful information as to the average treatment effect on the treated. It is clear that the interactive model needs a deeper methodological approach to prepare it for panel data besides just block-k cross-fitting and the methods we used for capturing fixed effects. Overall, while the naive and clustered interactive methods are flexible and general-purpose, they fall short in settings where structured panel data assumptions are crucial for identification and valid inference.

We also notice that the CRE approach generated the poorest results. This was an interesting observation as Fuhr and Papies [17] showed that the CRE model offered the most stable estimates in various data-generating processes. In our case, it is clear that CRE and CRE-

unit are the poorest performing methods for capturing fixed effects. Despite their findings, this approach consistently underperformed almost every other method. It is important to note that since we used the DoubleML package, we did not allow for different inputs of the outcome and treatment nuisance functions. Therefore, we omitted including the treatment mean, $\bar{D}$, in the covariate set, as it may lead to leakage since $\bar{D}$ would also be used in the treatment nuisance function estimation. Additionally, this setting is potentially at odds with the assumptions of the partially linear model, as the treatment is directly used as an input in the estimation phase. Nevertheless, we obtain estimates using CREs under the naive and clustered interactive models with the treatment mean included and still obtain poor results.

To investigate this further, we perform a diagnostic test to formally compare linear regression estimates obtained from a fixed effects model and those from the CRE model. The Hausman test allows us to evaluate the estimates in a linear regression setting between these two models. The null hypothesis states that both the CRE and fixed effects are consistent and a rejection of the null indicates that the CRE model insufficiently captures the unit-level unobserved heterogeneity. After running the test, we obtain p-values of 0.1013 and 0.0114 for the CRE and CRE-unit methods respectively. One can see that the CRE-unit is barely significant in sufficiently capturing the heterogeneity. The poorer performance of the CRE-unit may stem from its limited flexibility by including only unit-level means of the covariates. Although this test shows that the full CRE is viable in the linear setting, suggesting that the CRE approach is theoretically consistent for our data, this consistency does not guarantee strong performance with machine learning.

Additionally, we carefully examine the first-differencing method, as highlighted by Clarke and Polselli [13], who demonstrate in their simulation study that first-differencing yields the most reliable estimates in both linear and non-linear data-generating processes. Given these findings, it seems that the first-differencing method may offer a more consistent framework

for estimating treatment effects, particularly when dealing with potential endogeneity or time-varying effects that other methods, like CRE, may struggle to address for our particular setting.

We find that the estimates from the adjusted partially linear model are the most robust and statistically significant. Estimates obtained via the first-difference and demean methods uphold the original analysis by Chronopoulos et al. [12], showing a significant increase in the log-change of *Risk-Weighted Assets* for banks above $50 billion in assets after quarter 1 of 2018. Furthermore, our results show that the magnitude of the valid DML estimates are close to that of the original linear TWFE model, showing that the original data may not have strong non-linearity to justify the use of machine learning. This goes to show that DML can still uphold and strengthen the results obtained via traditional methods and is especially useful in cases when there is reason to believe the data is non-linear. Under the right conditions, it can help to relax functional form assumptions and is a valuable method of inference for panel data.

We recommend that researchers who are looking to use DML with panel data ensure that they have balanced panel data. Due to the inconsistent estimates we obtained from the varying approaches, regardless of whether we use the naive or adjusted method, we suggest that researchers take full advantage of different techniques for incorporating fixed effects in panel data. They should run DML on a diverse set of fixed effects techniques to ensure the most stable results. Even traditional tests such as the Hausman test, might ensure theoretical consistency of estimates in a linear setting, but transformations could still be ill suited for causal machine learning on the data.

Even though we utilize this method within the domain of financial regulatory analysis, DML can be implemented in a wide variety of settings that involve observing data and potential controls before and after an intervention. The assumption of unconfoundedness, which states

that, conditional on observed covariates, the treatment assignment is as good as random, is critical. The novel method for panel data from Clarke and Polselli [13] can be implemented in empirical settings where the data consists of multiple units observed over multiple time periods. For instance, studies ranging from social sciences to healthcare that evaluate policies affecting schools, hospitals, or municipalities can benefit from the combination of machine learning capabilities and causal inference that DML provides. This holds as long as there is a rich set of controls that appropriately satisfy the assumptions.

Future research could provide important methodological advances for causal inference in panel data settings and be an instrumental tool for policy researchers looking for stronger causal frameworks. One potential area is the theoretical development of the interactive model framework for panel data, which would extend current DML methods to estimate treatment effects. Due to the more generalized form of the interactive model and its ability to provide an estimate of the treated units (ATTE) under heterogeneous treatment effects, formally adapting the model to a panel setting would be very helpful for policy researchers working with similar data.

Another avenue for future work relating to the methodology of DML involves enhancing the design of Monte Carlo simulation studies in causal machine learning research. Much of the existing literature relies on data generating processes with Gaussian assumptions. Developing simulation designs that incorporate non-Gaussian and nonlinear functional forms could help researchers better assess the robustness of panel DML estimators and the effectiveness of panel-specific transformations. This would help to strengthen its applications and the external validity of simulation-based evaluations.

Chapter 7

Conclusions

The 2008 financial crisis was a significant macroeconomic event that led to the establishment of the Dodd-Frank Act, a landmark legislation that imposed stronger regulations on banks and allowed for greater government oversight. In 2018, the passage of the EGRRCPA rolled back some major provisions of the Dodd-Frank Act. This study reexamined the rollback of one such provision that relaxed regulations on banks with between $50 and $250 billion in assets, and analyzed the effect on bank risk after the intervention.

Since most financial policy analysis and bank risk literature relies on traditional causal inference methods that rely on strong linear assumptions, we sought to apply advanced techniques from causal machine learning. In doing so, we contribute to the sparse literature that uses causal machine learning for policy analysis and to the broader, growing body of work on DML.

We also add to the limited empirical research applying DML to panel data, as much of the existing work focuses on cross-sectional settings. Among the few studies that apply DML to panel data, most do not disclose the specific adjustments made to account for the structure of panel data. To address this gap, we evaluated developments in the DML literature and implemented the recent method proposed by Clarke and Polselli [13].

We demonstrated the viability of this approach in an empirical setting focused on financial policy analysis of the EGRRCPA and compared the results to strong baseline regression esti-

mates previously obtained. In addition to replicating the original findings and applying the DML framework, we show that simply using naive methods of DML on this data yields unreliable estimates. To ensure robust implementation, we employed strategies including data transformations, hyperparameter tuning, and cross-fitting to recover results that accurately reflect the advantages of DML in panel data contexts.

Through this analysis, we uphold the original findings showing that the rollback of a provision within the EGRRCPA, which raised the threshold for enhanced liquidity and capital standards from banks with $50 billion in assets to those with $250 billion, led to increased risk.

Due to the nature of machine learning algorithms, these methods are often viewed as black box approaches. However, this perception is not entirely accurate in the context of causal machine learning frameworks such as DML, which explicitly estimates the treatment effect from the nuisance components allowing for improved interpretability and robustness. Nevertheless, progress is still needed, particularly in adapting these methods to complex settings like panel data, before their use becomes standard practice in applied policy research. Additionally, the appeal and value of DML become especially clear in settings with high-dimensional or non-linear data.

Bibliography

[1] Alberto Alesina, Paola Giuliano, and Nathan Nunn. On the origins of gender roles: Women and the plough. *The quarterly journal of economics*, 128(2):469–530, 2013.

[2] Philipp Bach, Victor Chernozhukov, Malte S Kurz, and Martin Spindler. Doubleml-an object-oriented implementation of double machine learning in python. *Journal of Machine Learning Research*, 23(53):1–6, 2022.

[3] Philipp Bach, Oliver Schacht, Victor Chernozhukov, Sven Klaassen, and Martin Spindler. Hyperparameter tuning for causal inference with double machine learning: A simulation study. In *Causal Learning and Reasoning*, pages 1065–1117. PMLR, 2024.

[4] Anna Baiardi and Andrea A Naghi. The effect of plough agriculture on gender roles: A machine learning approach. *Journal of Applied Econometrics*, 2024.

[5] Anna Baiardi and Andrea A Naghi. The value added of machine learning to causal inference: Evidence from revisited studies. *The Econometrics Journal*, page utae004, 2024.

[6] Lakshmi Balasubramanyan, Naveen D Daniel, Joseph G Haubrich, and Lalitha Naveen. Impact of risk oversight functions on bank risk: Evidence from the dodd-frank act. *Journal of Banking & Finance*, 158:107049, 2024.

[7] Jack Bao, Maureen O'Hara, and Xing Alex Zhou. The volcker rule and corporate bond market making in times of stress. *Journal of Financial Economics*, 130(1):95–113, 2018.

[8] Zongxuan Chai, Tingting Zheng, et al. The explaimability of double machine learning

causal inference in quasi-natural experiments—a study based on county panel sample data. *Automation and Machine Learning*, 4(3):49–54, 2023.

[9] Neng-Chieh Chang. Double/debiased machine learning for difference-in-differences models. *The Econometrics Journal*, 23(2):177–191, 2020.

[10] Victor Chernozhukov, Denis Chetverikov, Mert Demirer, Esther Duflo, Christian Hansen, Whitney Newey, and James Robins. Double/debiased machine learning for treatment and structural parameters, 2018.

[11] Caroline Chin et al. *Measuring the Impact of Elections on Judge Behavior Using Machine Learning and Economics Tools*. PhD thesis, Massachusetts Institute of Technology, 2021.

[12] Dimitris K Chronopoulos, John OS Wilson, and Muhammed H Yilmaz. Regulatory oversight and bank risk. *Journal of Financial Stability*, 64:101105, 2023.

[13] Paul S Clarke and Annalivia Polselli. Double machine learning for static panel models with fixed effects. *The Econometrics Journal*, page utaf011, 2025.

[14] Stefano DellaVigna and Ethan Kaplan. The political impact of media bias. *Information and Public Choice*, page 79, 2008.

[15] Simeon Djankov, Tim Ganser, Caralee McLiesh, Rita Ramalho, and Andrei Shleifer. The effect of corporate taxes on investment and entrepreneurship. *American Economic Journal: Macroeconomics*, 2(3):31–64, 2010.

[16] Paul B Ellickson, Wreetabrata Kar, and James C Reeder III. Estimating marketing component effects: Double machine learning from targeted digital promotions. *Marketing Science*, 42(4):704–728, 2023.

[17] Jonathan Fuhr and Dominik Papies. Double machine learning meets panel data–promises, pitfalls, and potential solutions. *arXiv preprint arXiv:2409.01266*, 2024.

[18] Ziwang Gao, Lihui Cai, and Xiaolu Zhang. New industrial land use policy and firms' green technology innovation in china—an empirical study based on double machine learning model. *Frontiers in Environmental Science*, 12:1356291, 2024.

[19] Jacob H Hansen and Mathias V Siggaard. Double machine learning: Explaining the post-earnings announcement drift. *Journal of Financial and Quantitative Analysis*, pages 1–28, 2023.

[20] Karel Janda and Oleg Kravtsov. Regulatory stress tests and bank responses: Heterogeneous treatment effect in dynamic settings. *International Journal of Central Banking*, 18(2):1–49, 2022.

[21] Yuchen Jiang, Lei Li, and Yue Xu. Can digital economy improve urban ecological development? evidence based on double machine learning analysis. *Frontiers in Environmental Science*, 13:1542363, 2025.

[22] Kevin Nooree Kim and Ani L Katchova. Impact of the basel iii bank regulation on us agricultural lending. *Agricultural Finance Review*, 80(3):321–337, 2020.

[23] Michael C Knaus. A double machine learning approach to estimate the effects of musical practice on student's skills. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 184(1):282–300, 2021.

[24] Anoop Kumar, Suresh Dodda, Navin Kamuni, and Rajeev Kumar Arora. Unveiling the impact of macroeconomic policies: A double machine learning approach to analyzing interest rate effects on financial markets. *arXiv preprint arXiv:2404.07225*, 2024.

[25] Sijia Li, Wei Pan, Paul Siu Fai Yip, Jing Wang, Wenwei Zhou, and Tingshao Zhu. Uncovering the heterogeneous effects of depression on suicide risk conditioned by linguistic

features: A double machine learning approach. *Computers in Human Behavior*, 152: 108080, 2024.

[26] Prashant Loyalka, Anna Popova, Guirong Li, and Zhaolei Shi. Does teacher training actually work? evidence from a large-scale randomized evaluation of a national teacher training program. *American Economic Journal: Applied Economics*, 11(3):128–154, 2019.

[27] Nathan Nunn and Daniel Trefler. The structure of tariffs and long-term growth. *American Economic Journal: Macroeconomics*, 2(4):158–194, 2010.

[28] Dwayne Powell. Quantitative analysis of the impact of the economic growth, regulatory relief, and consumer protection act of 2018 on the cost of regulatory burden on community banks. *Journal of Accounting and Finance*, 22(4), 2022.

[29] Xuqing Wang, Yahang Liu, Guoyou Qin, and Yongfu Yu. Robust double machine learning model with application to omics data. *BMC bioinformatics*, 25(1):355, 2024.

[30] Jui-Chung Yang, Hui-Ching Chuang, and Chung-Ming Kuan. Double machine learning with gradient boosting and its application to the big n audit quality effect. *Journal of Econometrics*, 216(1):268–283, 2020.

[31] Jie Yuan and Shucheng Liu. A double machine learning model for measuring the impact of the made in china 2025 strategy on green economic growth. *Scientific Reports*, 14 (1):12026, 2024.

Appendices

Appendix A

First Appendix

Table A.1: Main - Partially Linear Results

	Pooled	Dummy	Demean	CRE Unit	CRE	First Difference
Linear Baseline	*0.0011 (0.0040)*	*0.0193 (0.0045)****	*0.0193 (0.0044)****	*0.0184 (0.0037)****	*0.0194 (0.0044)****	*0.0189 (0.0054)****
Random Forest	0.0026 (0.0040)	0.0042 (0.0037)	0.0199 (0.0044)***	0.0074 (0.0039)	0.0029 (0.0083)	0.0231 (0.0074)***
Random Forest Base	0.0112 (0.0057)**	0.0113 (0.0057)**	0.0222 (0.0049)***	0.0211 (0.0067)***	0.0311 (0.0155)**	0.0233 (0.0084)***
Gradient Boost	0.0037 (0.0044)	0.0025 (0.0093)	0.0193 (0.0044)***	0.0117 (0.0047)**	0.0199 (0.0227)	0.0197 (0.0059)***
Gradient Boost Base	0.0083 (0.0061)	0.0154 (0.0097)	0.0221 (0.0051)***	0.0148 (0.0076)*	0.1011 (0.0723)	0.0252 (0.0095)***
Logit/OLS	0.0004 (0.0033)	0.0246 (0.0066)***	0.0194 (0.0045)***	0.0041 (0.0032)	0.0036 (0.0033)	0.0196 (0.0056)***
Ridge	0.0012 (0.0033)	0.0192 (0.0044)***	0.0194 (0.0044)***	0.0044 (0.0034)	0.0046 (0.0036)	0.0195 (0.0056)***
Lasso	0.0007 (0.0032)	0.0106 (0.0040)***	0.0189 (0.0044)***	0.0039 (0.0031)	0.0042 (0.0031)	0.0190 (0.0055)***
Observations	1820	1820	1820	1820	1820	1729

Uses the orthogonal score functions from the partially linear setting. Significance levels indicated by '***', '**', '*' for the 1, 5, and 10 percent levels respectively.

Table A.2: Clarke & Polselli - Partially Linear Results

	Demean	CRE Unit	CRE	First Difference
Linear Baseline	*0.0193 (0.0044)****	*0.0184 (0.0037)****	*0.0194 (0.0044)****	*0.0189 (0.0054)****
Random Forest Tune	0.0193 (0.0041) ***	0.0147 (0.0041) ***	-0.0079 (0.0131)	0.0142 (0.0060) **
Random Forest Default	0.0192 (0.0042) ***	0.0145 (0.0042) ***	-0.0060 (0.0129)	0.0144 (0.0062) **
Gradient Boost Tune	0.0171 (0.0039) ***	0.0089 (0.0045) **	0.1994 (0.2673)	0.0134 (0.0050) ***
Gradient Boost Default	0.0147 (0.0043) ***	0.0084 (0.0045) *	-0.2442 (0.5799)	0.0162 (0.0071) **
Logit/OLS	0.0189 (0.0044) ***	0.0178 (0.0038) ***	0.0183 (0.0044) ***	0.0164 (0.0057) ***
Lasso	0.0192 (0.0040) ***	0.0074 (0.0056)	0.0099 (0.0089) ***	0.0173 (0.0065) ***
Observations	1820	1820	1820	1729

Uses the orthogonal score functions from the partially linear setting. Significance levels indicated by '***', '**', '*' for the 1, 5, and 10 percent levels respectively.

Table A.3: Main - Interactive Model Results

		Pooled	Dummy	CRE	CRE Unit	First Difference
Linear Baseline		*0.0011 (0.0040)*	*0.0193 (0.0045)****	*0.0194 (0.0044)****	*0.0184 (0.0037)****	*0.0189 (0.0054)****
Random Forest Tuned	**ATE:**	-0.0039 (0.0027)	-0.0046 (0.0028)	-0.0030 (0.0012)**	-0.0032 (0.0019)	0.0115 (0.0037)***
	ATTE:	0.0015 (0.0031)	0.0045 (0.0033)	0.0017 (0.0028)	0.0065 (0.0030) **	0.0241 (0.0079)***
Random Forest Base	**ATE:**	0.0108 (0.0096)	-0.0054 (0.0021)***	0.0014 (0.0012)	0.0023 (0.0015)	0.0129 (0.0044)***
	ATTE:	0.0014 (0.0038)	0.0058 (0.0031)	0.0021 (0.0028)	0.0091 (0.0033) ***	0.0228 (0.0098)**
Gradient Boost Tuned	**ATE:**	-0.0016 (0.0049)	-0.0044 (0.0022)	-0.0040 (0.0011)***	-0.0019 (0.0018)	0.0124 (0.0045)***
	ATTE:	0.0005 (0.0033)	-0.0031 (0.0041)	0.0025 (0.0028)	0.0086 (0.0030) ***	0.0192 (0.0070)***
Gradient Boost Base	**ATE:**	0.0119 (0.0171)	0.0124 (0.0156)	0.0064 (0.0013)***	0.0223 (0.0159)	0.0116 (0.0077)
	ATTE:	0.0011 (0.0219)	0.0032 (0.0060)	-0.0023 (0.0033)	0.0147 (0.0113)	0.0264 (0.0108)**
Logit/OLS	**ATE:**	-0.0048 (0.0039)	-0.0068 (0.0012)***	0.0184 (0.0039)***	0.0053 (0.0044)	0.0041 (0.0057)
	ATTE:	0.0002 (0.0037)	0.0183 (0.0031)***	0.0044 (0.0035)	0.0042 (0.0035)	0.0196 (0.0056)***
Ridge	**ATE:**	-0.0048 (0.0039)	-0.0031 (0.0012) ***	-0.0032 (0.0038)	-0.0024 (0.0043)	0.0123 (0.0049)**
	ATTE:	0.00004 (0.0037)	0.0177 (0.0031)***	0.0043 (0.0035)	0.0039 (0.0035)	0.0194 (0.0055)***
Lasso	**ATE:**	-0.0048 (0.0039)	-0.0059 (0.0012)***	-0.0033 (0.0038)	-0.0036 (0.0042)	0.0102 (0.0052)*
	ATTE:	0.0004 (0.0036)	0.0006 (0.0029)	0.0042 (0.0035)	0.0038 (0.0035)	0.0189 (0.0055)***
Observations		**1820**	**1820**	**1820**	**1820**	**1729**

Main results from the Interactive model. Uses ATE and ATTE estimates. Significance levels indicated by '***', '**', '*' for the 1, 5, and 10 percent levels respectively.

Table A.4: Main - Clustered Interactive Model Results

		CRE	CRE Unit	First Difference
Linear Baseline		*0.0194 (0.0044)****	*0.0184 (0.0037)****	*0.0189 (0.0054)****
Random Forest Tuned	**ATE:**	-0.0051 (0.0026)**	-0.0052 (0.0037)	0.0123 (0.0033)***
	ATTE:	-0.0024 (0.0041)	0.0040 (0.0034)	0.0143 (0.0058)**
Random Forest Base	**ATE:**	-0.0019 (0.0036)	-0.0036 (0.0088)	0.0125 (0.0033)***
	ATTE:	-0.0050 (0.0053)	0.0085 (0.0051)	0.0122 (0.0077)
Gradient Boost Tuned	**ATE:**	-0.0062 (0.0086)	-0.0021 (0.0076)	0.0114 (0.0046)**
	ATTE:	-0.0157 (0.0149)	0.0047 (0.0038)	0.0153 (0.0071)**
Gradient Boost Base	**ATE:**	-0.0267 (0.0313)	-0.0362 (0.0495)	0.0083 (0.0062)
	ATTE:	-0.0075 (0.0288)	0.0297 (0.0558)	0.0158 (0.0086)*
Logit/OLS	**ATE:**	0.0183 (0.0045)***	0.0062 (0.0049)	0.0029 (0.0103)
	ATTE:	0.0051 (0.0045)	0.0059 (0.0043)	0.0195 (0.0056)***
Ridge	**ATE:**	-0.0028 (0.0043)	-0.0011 (0.0051)	0.0123 (0.0041)***
	ATTE:	0.0044 (0.0044)	0.0035 (0.0055)	0.0198 (0.0056)***
Lasso	**ATE:**	-0.0029 (0.0041)	-0.0033 (0.0046)	0.0118 (0.0042)***
	ATTE:	0.0039 (0.0047)	0.0061 (0.0051)	0.0182 (0.0055)***
Observations		**1820**	**1820**	**1729**

Main results from the interactive model using *block-k* cross-fitting. Uses ATE and ATTE estimates. Significance levels indicated by '***', '**', '*' for the 1, 5, and 10 percent levels respectively.

Appendix B

Second Appendix

Deriving the Interactive Model ATE Estimator from the Neyman Orthogonal Score Function

To identify the parameter θ, solve the population moment condition:

$$\mathbb{E}[\psi(W;\theta,\eta)] = 0.$$

In the DoubleML library [2] the general orthogonal score function is defined as:

$$\begin{aligned}\psi(W;\theta,\eta) &:= \omega(Y,D,X)\cdot(g(1,X)-g(0,X)) \\ &\quad + \bar{\omega}(X)\cdot\left(\frac{D(Y-g(1,X))}{m(X)} - \frac{(1-D)(Y-g(0,X))}{1-m(X)}\right) - \theta \\ &= \psi_b(W;\eta) + \psi_a(W;\eta)\theta,\end{aligned}$$

where $\omega(Y,D,X)$ and $\bar{\omega}(X)$ are user-defined weights.

Here, the score is linear in θ with:

$$\psi_a(W;\eta) = -1,$$

and:

$$\psi_b(W;\eta) = \omega(Y,D,X)\cdot(g(1,X)-g(0,X)) \\ + \bar{\omega}(X)\cdot\left(\frac{D(Y-g(1,X))}{m(X)} - \frac{(1-D)(Y-g(0,X))}{1-m(X)}\right).$$

To estimate the ATE, both weights are set to one:

$$\omega(Y,D,X) = 1, \quad \bar{\omega}(X) = 1.$$

Substituting into the score yields:

$$\psi_b(W;\eta) = g(1,X) - g(0,X) + \frac{D(Y-g(1,X))}{m(X)} - \frac{(1-D)(Y-g(0,X))}{1-m(X)}.$$

Then, solving the moment condition:

$$\mathbb{E}[\psi(W;\theta,\eta)] = \mathbb{E}[\psi_b(W;\eta)] - \theta = 0 \quad \Rightarrow \quad \boxed{\theta = \mathbb{E}[\psi_b(W;\eta)].}$$

Deriving the Interactive Model ATTE Estimator from the Neyman Orthogonal Score Function

To identify the parameter θ, solve the moment condition:

$$\mathbb{E}[\psi(W;\theta,\eta)] = 0.$$

In the ATTE case, the weights in the general score function are specified as:

$$\bar{\omega}(X) = \frac{m(X)}{\mathbb{E}[D]}, \quad \omega(Y, D, X) = \frac{D}{\mathbb{E}[D]}.$$

Additionally, we also assume that $g(1, X) = Y$, since we are only working with the treated sample.

Substituting these components into the general DoubleML score, and using the fact that $\psi_a(W; \eta) = -1$, the $\psi_b(W; \eta)$ term becomes:

$$\psi_b(W; \eta) = \frac{D}{\mathbb{E}_n[D]} (Y - g(0, X)) - \frac{m(X)}{\mathbb{E}_n[D]} \cdot \left(\frac{(1 - D)(Y - g(0, X))}{1 - m(X)} \right).$$

Then, setting the expectation of the full score function equal to zero:

$$\mathbb{E}[\psi_a(W; \eta) \cdot \theta + \psi_b(W; \eta)] = 0,$$

and recalling that $\psi_a(W; \eta) = -1$, set θ as:

$$\boxed{\theta = \mathbb{E}[\psi_b(W; \eta)].}$$

The score function 4.11 from Chernozhukov et al. [10] provides us the same result as the general score. The orthogonal score function for the ATTE as defined in Chernozhukov et al. [10]:

$$\psi(W; \theta, \eta) = \frac{D(Y - \bar{g}(X))}{p} - \frac{m(X)(1 - D)(Y - \bar{g}(X))}{p(1 - m(X))} - \frac{D}{p} \cdot \theta,$$

where: $p = \mathbb{E}[D]$ is the treatment probability and $\bar{g}(X) = \mathbb{E}[Y|D = 0, X]$ is the outcome model for the untreated.

Rewriting this:

$$\psi(W;\theta,\eta) = \psi_b(W;\eta) - \frac{D}{p}\cdot\theta,$$

where:

$$\psi_b(W;\eta) = \frac{D(Y-\bar{g}(X))}{p} - \frac{m(X)(1-D)(Y-\bar{g}(X))}{p(1-m(X))}.$$

Taking expectations on both sides and solving the moment condition:

$$\mathbb{E}[\psi(W;\theta,\eta)] = 0 \quad \Rightarrow \quad \mathbb{E}[\psi_b(W)] = \mathbb{E}\left[\frac{D}{p}\right]\cdot\theta.$$

Since $\mathbb{E}[D/p] = 1$, it follows that:

$$\boxed{\theta = \mathbb{E}[\psi_b(W)].}$$

This is algebraically equivalent to the result obtained from the general DoubleML score after simplification under ATTE-specific assumptions.

Deriving the ATE Estimator from the Partialling-Out Score Function

In the partially linear model of Chernozhukov et al. [10], the regression equation is:

$$Y = D\theta_0 + g(X) + U, \quad \text{with } \mathbb{E}[U|X,D] = 0.$$

To estimate θ_0, we use the Robinson-style orthogonal score function:

$$\psi(W;\theta,\eta) := \{Y - \ell(X) - \theta(D - m(X))\}\,(D - m(X)), \quad \text{where } \eta = (\ell, m).$$

Here, $\ell(X) = \mathbb{E}[Y|X]$ is the outcome model and $m(X) = \mathbb{E}[D|X]$ is the propensity score. This score is Neyman orthogonal and robust to regularization bias in both nuisance components.

The moment condition $\mathbb{E}[\psi(W; \theta, \eta)] = 0$ leads to the estimator:

$$\boxed{\theta = \frac{\mathbb{E}[(Y - \ell(X))(D - m(X))]}{\mathbb{E}[(D - m(X))^2]}.}$$

In empirical implementation, the expectations are replaced by sample averages, and the nuisance functions are learned via machine learning with cross-fitting.