Erasmus University Rotterdam
Erasmus School of Economics
Bachelor Thesis Econometrics

Application of Machine Learning in Causal Inference: Analyzing the Effects of Abortion Legalization on Crime Rates

Sterre Bras (563676)

Supervisor:	S.J. Koobs
Second assessor:	J. Durieux
Date final version:	1st July 2024

Abstract

High-dimensional data is increasingly common in empirical economics, posing significant challenges for causal inference. This thesis contributes to the ongoing debate on the abortion and crime hypothesis by exploring post-double-selection and double machine learning (DML), specifically employing the least absolute shrinkage and selection operator (Lasso) and extreme gradient boosting (XGBoost). The results indicate that Lasso-based post-double-selection and XGBoost-based DML outperform Lasso-based DML in predictive accuracy. Both methods suggest a significant negative effect of abortion rates on violent- and property crime rates for most time periods, though the estimated effects are smaller than those reported by Donohue III and Levitt (2001) and Donohue and Levitt (2020). No significant negative effect is found regarding the murder crime rate.

Keywords: Double machine learning, Post-double-selection, XGBoost, Abortion- and crime rates

1 Introduction

High-dimensional datasets, which involve a large number of variables relative to the number of observations, are becoming more prevalent in empirical economics. Numerous statistical methods are at hand for building prediction models when dealing with high-dimensional data. Yet, the aim of reaching high predictive accuracy differs fundamentally from the aim of accurate effect estimation (Shmueli, 2010). Consequently, while performing well at prediction tasks, these statistical methods tend to yield incorrect conclusions when it comes to inference regarding regression coefficients (Leeb & Pötscher, 2008). Nonetheless, estimating causal effects is of interest across a wide range of disciplines: Labor economists are interested in the impact of education on wages (Card, 1999), marketing managers aim to understand the influence of price changes on demand (Bijmolt, Van Heerde & Pieters, 2005), and social scientists study the effects of emergencies, such as the COVID-19 pandemic, on prosocial behaviour (Bilancini et al., 2022).

Despite the availability and occurrence of high-dimensional datasets, the inclusion of an excessive number of regressors relative to sample size presents notable challenges, leading to poor properties when looking at standard econometric methods such as least squares: capturing both the informative relationships between predictor and outcome variables as well as the noise embedded in the sample, limiting its utility for forming predictions out-of-sample (Belloni, Chernozhukov & Hansen, 2014a). Standard econometric methods, such as least squares, even become infeasible when the number of variables surpasses the number of observations. Building upon the preceding, Belloni et al. (2014a) state that the key concept of analyzing high-dimensional data is the necessity of dimension reduction for extracting meaningful conclusions. Accordingly, sparsity is a key condition that makes inference in high-dimensional data environments possible (Belloni, Chernozhukov & Hansen, 2014b). I.e., it is assumed there is a small enough set of variables, within a broad set of potential variables, that can approximate the treatment effect adequately. Given the vast amount of available data, researchers often face the challenge of selecting appropriate controls. Economic intuition can point to a selection of important variables but fails to clearly specify which ones are critical or in what functional form the variables should enter the model. Relying solely on a limited set of variables assumes that there are no

additional omitted variables, higher order terms and interaction terms, that are related to both the treatment variable and the outcome variable. Therefore, flexibly controlling for a large set of variables, encompassing higher order terms and interaction terms, is desired to enhance the credibility of inference on the causal effect of the treatment variable (Belloni et al., 2014b). To account for the realistic scenario where a researcher cannot know beforehand which regressors to include, methods that conduct data-driven variable selection are sought after and of growing interest.

A common progression in addressing this challenge is to explore ML algorithms, as they have proved to be a powerful tool for predicting outcomes in complex, nonlinear settings, and for their ability to deal with high dimensional data (e.g. Hastie, Tibshirani and Friedman (2009)). Moreover, several ML methods, such as regularization- and tree-based methods, perform data-driven variable selection by default (Fuhr, Berens & Papies, 2024). ML algorithms trade higher bias for reduced variance (e.g. via regularization) (Shi, Mao, Yang & Li, 2023), making inference invalid (Athey et al., 2018). Furthermore, when the same dataset is utilized for both training the model and assessing its performance, the ML estimator is prone to overfitting, which results in additional bias (Fuhr et al., 2024). Directly employing ML for causal inference would therefore be a naive approach (Belloni et al., 2014b).

Nevertheless, even with data-driven variable selection methods, it is important to note that realistically, variable selection mistakes are bound to happen, and that estimation approaches must be robust for the so-called model selection mistakes. A naive approach could be to estimate and do inference by only looking at which variables are important in learning the outcome variable. Any variable with strong correlation to the treatment variable and a moderate effect on the outcome will be generally be omitted, as such a variable provides little additional predictive value for the outcome when the treatment variable is already included. As a result, this approach is prone to lead to a substantial omitted-variables bias. As such single equation approaches hinge on this perfect model selection assumption, Belloni et al. (2014b) propose a double-selection procedure: using two data-driven variable selection steps to select a set of variables valuable for predicting the outcome variable and a set of variables valuable for predicting the treatment variable. Thereby contributing to ensuring the validity of post-model-selection inference by identifying potentially significant confounders, i.e. variables that affect both the outcome and treatment variable of interest.

Expanding upon this insight and the possible use of flexible ML methods for causal inference, Chernozhukov et al. (2018) introduce two key elements in their double machine learning (DML) method to protect DML from the bias introduced via regularization and the possible risk of overfitting respectively. The first key element involves Neyman-orthogonal moments, which are obtained by partialling out the effect of the controls on the treatment variable before predicting the outcome variable using those same controls. The second key element is cross-fitting, which prevents overfitting as a result of utilizing the same data for training and assessing performance of the method. Accordingly, DML enables the utilization of diverse flexible ML methods for the variable selection tasks while still obtaining approximately unbiased estimates, even when faced with complex functional forms and many confounders (Chernozhukov et al., 2018). Despite the primary advantage of DML being able to employ diverse flexible ML methods for inference,

and thereby its capability to address nonlinear confounding without requiring knowledge of the underlying functional form, researches commonly employ Lasso within the DML framework (Fuhr et al., 2024). Nevertheless, Fuhr et al. (2024) employed Lasso-based DML (DML Lasso) in several simulation settings, and conclude that it does not provide additional flexibility compared to a classical OLS regression in fitting complex nonlinear relationships when only considering raw variables without transformations. Common practice for modeling nonlinear relationships when dealing with methods that cannot inherently handle them, is to include a variety of transformations, resulting in a large set of potential controls. However, researches often lack guidance on questions such as which variables should interact and to what extent, the adequate polynomial order, and other potentially useful functional forms. In addition, as the set of potential controls increases to handle these nonlinearities, the performance of DML Lasso deteriorates (Fuhr et al., 2024).

Consequently, this thesis extends its focus beyond DML Lasso by employing extreme gradient boosting (XGBoost). Chen and Guestrin (2016) introduce XGBoost, an advanced ensemble machine learning algorithm that combines decision trees with gradient boosting. XGBoost is able to inherently handle nonlinear and complex relationships (Chen & Guestrin, 2016), eliminating the need to manually include transformations and the associated risk of omitting relevant ones. Across numerous machine learning and data mining challenges, XGBoost has achieved widespread recognition for its performance (Chen & Guestrin, 2016). Specifically, XGBoost has been recognized as a powerful ML method to use within the DML framework (Shi et al., 2023). Moreover, Fuhr et al. (2024) propose DML XGBoost as the preferred method across various contexts, even advocating for this method to be adopted as the standard approach

A commonly studied subject in causal inference is the cause behind the large, widespread, and persistent decline in crime rates since 1991 in the United States (U.S.). Donohue III and Levitt (2001) present evidence suggesting that the legalization of abortion has been a significant factor contributing to this decline, possibly explaining up to 50% of the crime reduction observed since 1991. Furthermore, Donohue III and Levitt (2001) predicted that legalized abortion would lead to persistent 1% decline annually over the next two decades (from 1997). This abortion and crime hypothesis is primarily based on two elements: First, economically disadvantaged individuals, such as teenagers, are more likely to seek abortions (Levine, Staiger, Kane & Zimmerman, 1999) and children born to these mothers are more likely to engage in crime during adolescence (Comanor & Phillips, 1995). Secondly, women may choose abortion to strategically plan the timing of childbearing, enabling them to delay childbearing when current conditions are suboptimal.

However, the findings of Donohue III and Levitt (2001) have generated considerable controversy and subsequent critical academic comments (Donohue & Levitt, 2020). For instance, Zimring et al. (2006) discuss that both sides of the debate are plausible but lack credible proof. Additionally, Joyce (2006) replicates the analyses of Donohue III and Levitt (2001) and their subsequent studies in 2004 and 2006, and finds statistically insignificant results when adjusting for serial correlation. Another point of critique is the so-called ad-hoc manner of variable selection, without taking into consideration other variables or higher order trends (Belloni et al., 2014b). Consequently, Belloni et al. (2014b) state that the conclusions drawn by Donohue III and Levitt (2001) depend strongly

on the assumption that there exist no additional unobserved state-level factors that correlate with abortion, the original controls, and the crime rates. On the other hand, there are also several extensions that align with the initial discoveries of Donohue III and Levitt (2001). For instance, Woody, Carvalho, Hahn and Murray (2020a) and Woody, Carvalho and Murray (2020b) provide evidence for a significant negative causal relationship consistent with Donohue III and Levitt (2001) using Bayesian tree ensembles. Furthermore, they argue that earlier replication studies may be afflicted by the phenomenon of a bias towards false-negative replication results. This phenomenon emerges from the flexibility researchers have in the setup and analysis when replicating a previously published finding, generally resulting in a bias favoring false-negative replication results (Bryan, Yeager & O'Brien, 2019). Furthermore, two decades after their original research, Donohue and Levitt (2020) re-examined their hypothesis using seventeen more years of data, employing the same methodologies as Donohue III and Levitt (2001). In fact, validating their earlier prediction of sustained crime decline, as a consequence of increase in abortion rates.

Concluding, there continues to be a significant diversity of the abortion and crime hypothesis among academics. This thesis adds to this debate by employing a post-double-selection method and two DML methods to this hypothesis. The dataset utilized in this thesis is obtained from the replication package of Donohue and Levitt (2020). The three dependent variables are violent-, property- and murder crime rates and with the effective abortion rate being the treatment variable, as defined by Donohue III and Levitt (2001), which averages state birth cohort abortion rates, weighted by each age group's share of national arrests for a particular crime type in 1985. The sample is split into two sub-sample periods, to accommodate a detailed analysis. This thesis examines a post-double-selection method using Lasso (post-double-Lasso), DML Lasso and DML XGBoost. For the methods involving Lasso, additional to the control variables from the replication package of Donohue and Levitt (2020), squared, cubed, and interaction terms of these controls are included in the dataset. The reliability of the models' treatment effect estimate is assessed by comparing their predictive accuracy. This approach is motivated by Fuhr et al. (2024) and Shi et al. (2023), which find that the predictive accuracy of ML methods within DML aligns with their ability to recover the true treatment effect. Nevertheless, this finding only holds under the strong condition that all relevant confounders are accounted for. In addition, a Monte-Carlo simulation example is conducted to further elaborate on the methods' performances.

This thesis finds that post-double-Lasso and DML XGBoost exhibit equal predictive accuracy and substantially outperform DML Lasso when evaluated on the abortion and crime hypothesis. Post-double-Lasso and DML XGBoost are therefore marked as more reliable compared to DML Lasso in this context. Consistent with Donohue III and Levitt (2001) and Donohue and Levitt (2020), this thesis demonstrates that increased abortion rates are associated with decreasing violent- and property crime rates across all three time periods. Nonetheless, not all three models agree on the significance of these negative effects for each time period. In addition, post-double-Lasso and DML XGBoost, the most reliable models, both yield effect estimates of substantial smaller magnitude than those found by Donohue III and Levitt (2001) and Donohue and Levitt (2020), indicating a smaller effect of abortion rates on violent- and property crime rate than they advocate for. For the effect of abortion rate on murder crime rate, none of my models find a significant negative effect. Post-double-Lasso and DML XGBoost even find positive effect

estimates for the period ranging from 1998-2014, making the hypothesis of Donohue III and Levitt (2001) and Donohue and Levitt (2020) regarding murder crime rate even more questionable.

Through an analysis of post-double-selection and the application of DML, an approach not previously used in this context, this thesis contributes to the ongoing debate on the abortion and crime hypothesis. Furthermore, this thesis examines not only the frequently discussed time period of 1985-1997 studied in Donohue III and Levitt (2001), but also the seventeen years of data that became available after the original publication in 2001, studied by Donohue and Levitt (2020). Enabling the investigation of whether the previously advocated effect still holds when studying another time period. Lastly, to my knowledge and to the knowledge of the extensive DML literature review performed by Fuhr et al. (2024), this thesis is one the few research efforts that employs the DML approach using XGBoost. Consequently, this enables a comparison between the commonly used Lasso method within DML against a more flexible ML method.

This thesis proceeds as follows. Section 2 describes the data used to perform inference on the abortion and crime hypothesis. Section 3 presents the methods and accesses and discusses their performance in a simulated setting. The results on the abortion and crime hypothesis are discussed in Section 4 and Section 5 concludes.

2 Data

The data utilized in this study is obtained from the replication package of Donohue and Levitt (2020). Following Belloni et al. (2014a) and for simplicity purposes, all observations for state Washington DC are deleted from the dataset. Accordingly, we do not include a population weight variable because the weighted results closely resemble the unweighted results excluding Washington DC (Donohue III & Levitt, 2001). The dataset encompasses a total of 1500 observations, covering the period from 1985 to 2014. To facilitate a detailed analysis and comparison, the dataset is split up into two periods, with the first sub-sample period being the period considered in Donohue III and Levitt (2001), spanning from 1985 to 1997 including 650 observations. The second sub-sample time period, spanning from 1998 to 2014 including 850 observations, represents the period containing data which became available after the publication of Donohue III and Levitt (2001). The dependent variables, representing crime rates for three different crime types, are retrieved from the FBI Uniform Crime Reporting Statistics (UCR). Specifically, the three dependent variables are designed as the natural logarithm of violent, property and murder crime rates. To standardize the crime rates across states and time, the crime rates are calculated per 1000 population using population data from FBI UCR.

The treatment variable, abortion rate, is computed as the "effective abortion rate", a metric introduced by Donohue III and Levitt (2001). The effective abortion rate averages state birth cohort abortion rates, weighted by each age group's share of 1985 national arrests for a particular crime type. The exact treatment variable thus differs across the three dependent variables. Following Donohue and Levitt (2020) and Donohue and Levitt (2004), I use abortion data by state of residence from the Alan Guttmacher Institute (AGI) and live birth data from NBER Vital Statistics Natality Birth Data (1970 - 1994) and CDC Wonder Natality Data (1995 - 2014).

Similar to Donohue III and Levitt (2001), the following eight state level control variables are considered: prisoners and police per capita, an indicator for concealed handgun laws presence,

lagged state welfare generosity, per capita beer consumption, per capita income, unemployment rate, and poverty rate. For a detailed description of these control variables and their data sources, this thesis refers to Appendix G of Donohue and Levitt (2020). Moreover, fixed effects for both state and time are accounted for by differencing all variables and including dummy variables respectively. This results in datasets of 39, 22, and 26 variables for the full sample period, first- and second sub-sample period respectively. To enhance the modeling of complex relationships for methods that cannot do so inherently, we include quadratic, cubic, and interaction terms. First, quadratic terms for all continuous variables are added. Then, I create interaction terms between all original variables and the quadratic terms. This approach allows for interactions between different variables as well as the original variables and their quadratic terms, creating cubed terms of the continuous variables. This results in datasets of 167, 150, and 154 potential controls for the full sample period, first- and second sub-sample period respectively.

3 Methodology

The following partially linear model specification is considered:

$$y_i = d_i'\alpha_0 + g(x_i) + \zeta_i \tag{1}$$

$$d_i = m(x_i) + \upsilon_i \tag{2}$$

where y_i represents the outcome variable, d_i represents the treatment variable for which we seek to estimate α_0, ζ_i and υ_i represent the error terms. The confounders x_i affect the treatment variable through $m(x_i)$ and the outcome variable through $g(x_i)$, which can be approximated by $m(x_i) = x_i'\eta_0 + r_{mi}$ and $g(x_i) = x_i'\beta_0 + r_{gi}$. Here x_i is the vector of controls, which could include transformations and interactions of the raw initial controls, r_{mi} and r_{gi} are corresponding approximation errors. The sparsity condition can now be outlined as the existence of approximations $x_i'\eta_0$ and $x_i'\beta_0$ that demand only a small number of non-zero coefficients to mitigate r_{mi} and r_{gi} small relative to the conjectured size of the estimation error. More formally,

$$\sqrt{\hat{E}(r_{gi}^2)} \lesssim \sqrt{s/n} \text{ and } \sqrt{\hat{E}(r_{mi}^2)} \lesssim \sqrt{s/n}. \tag{3}$$

3.1 Post-double selection

If one would only apply a variable selection method to Equation (1), the method is prone to omit variables with strong predictive power for d_i but less predictive power for y_i. Such omissions could lead to substantial omitted-variables bias (Belloni et al., 2014a). Belloni et al. (2014b) introduce an extra variable selection step applied to Equation (2) to overcome this problem. Their post-double selection method can be described by the following three steps:

1. Perform variable selection on Equation (2), thereby selecting a set of control variables, $\hat{I}_1$, that enhance accuracy on predicting the treatment d_i.

2. Perform variable selection on Equation (1), thereby selecting a set of control variables, $\hat{I}_2$, that enhance accuracy for predicting the outcome y_i.

3. Perform a linear regression of y_i on the treatment d_i and the selected variables $\hat{I} = \hat{I}_1 \cup \hat{I}_2$.

This approach produces the post-double-selection estimator $\hat{\alpha}$ of α_0:

$$(\hat{\alpha}, \hat{\beta}) = \underset{\alpha \in \mathbb{R}, \beta \in \mathbb{R}^p}{\operatorname{argmin}} \{\mathbb{E}_n[(y_i - d_i\alpha - x_i'\beta)^2] : \beta_j = 0, \forall j \notin \hat{I}\}. \tag{4}$$

For the variable selection tasks, this thesis employs the least absolute shrinkage and selection operator (Lasso) estimator, proposed by Tibshirani (1996). Lasso modifies the typical least squares objective by incorporating a penalty term and is established as the solution to the following optimization objective:

$$\min_{\beta \in \mathbb{R}^p} \mathbb{E}_n[(y_i - x_i'\beta)^2] + \frac{\lambda}{n}||\beta||_1 \tag{5}$$

where $||\beta||_1 = \sum_{j=1}^{p} |\beta_j|$. The penalty term $\frac{\lambda}{n}||\beta||_1$ induces sparsity in the coefficients by shrinking some of the coefficients to zero, thereby performing variable selection. Increasing λ intensifies the shrinkage effect, causing more coefficients to be reduced to zero and thus simplifying the model. Hence, parameter λ controls the trade-off between fitting the model well to the training data and reducing the model's complexity. For a detailed description about how Lasso minimizes the optimization problem in Equation (5), I refer to Tibshirani (1996). In high-dimensional data, Lasso offers substantial advantages by reducing model complexity and avoiding overfitting by selecting only the necessary predictors. In line with Belloni et al. (2014b), this thesis focuses on Lasso as a variable selection step within the post-double-selection framework. Despite this, it is important to note that this method and its properties are generalized to other variable selection methods under the assumption that the method is sparse and can yield reliable approximations for the functions g and m in Equations (1) and (2) respectively (Belloni et al., 2014b). For the remainder of this thesis, the post-double-selection model employing Lasso will be referred to as post-double-Lasso.

3.1.1 Monte Carlo Simulation

The improved inference performance of the post-double-Lasso estimator compared to single equation approaches can be demonstrated replicating the Monte Carlo simulation example conducted in Belloni, Chernozhukov and Hansen (2011). In this simulation example, the following simplified model is used:

$$y_i = d_i'\alpha_0 + \tilde{x}_i'\beta_0 + \zeta_i, \qquad \zeta_i \sim N(0,1) \tag{6}$$

$$d_i = \tilde{x}_i'\eta_0 + \upsilon_i, \qquad \upsilon_i \sim N(0,1) \tag{7}$$

where covariates $\tilde{x} \sim N(0, \Sigma)$ and $\Sigma_{kj} = (0.5)^{|j-k|}$. Sample size n is 100, the number of covariates in x is 200 and α_0 is set to equal 1. Furthermore,

$$\beta_0 = (1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, 0, 0, 0, 0, 0, 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, 0, ..., 0)'$$
$$\eta_0 = (1, 1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \frac{1}{6}, \frac{1}{7}, \frac{1}{8}, \frac{1}{9}, \frac{1}{10}, 0, ..., 0)'.$$

The studied estimation strategies are Lasso, post-Lasso, indirect-post-Lasso, post-double-Lasso, double selection Oracle, and Oracle. Post-Lasso represents the procedure where Lasso performs variable selection on Equation (6), whereafter ordinary least squares (OLS) is applied to the dataset including only the selected variables. Similarly, indirect-post-Lasso represents the procedure where Lasso performs variable selection on Equation (7), whereafter OLS is applied to the dataset including only the selected variables. For all Lasso variable selection steps, λ is set according to the following X-dependent penalty level proposed by Belloni and Chernozhukov (2013):

$$\lambda = 2c\hat{\sigma}\Lambda(1-\gamma|X) \tag{8}$$

$$\Lambda(1-\gamma|X) = (1-\gamma) - \text{quantile of n}||\mathbb{E}_n[x_i h_i]||_\infty \quad |X,$$

where $X = [x_1, ..., x_n]'$, h_i are i.i.d. N(0,1) $c = 1.1$ and $\gamma = 0.05$. In practice, λ in Equation (8) often leads to overpenalization (Belloni et al., 2011). To prevent overpenalization, I utilize the iteration procedure, described in (Belloni & Chernozhukov, 2011) to estimate σ. This procedure is described in Appendix A.1. Finally, the Oracle estimates α_0 by applying OLS of $y_i - \tilde{x}_i'\beta_0$ on d_i using knowledge of the true coefficient structure β_0. Double selection Oracle estimates α_0 by applying OLS of $y_i - \tilde{x}_i'\beta_0$ on $d_i - \tilde{x}_i'\eta_0$ using knowledge of the true coefficient structures β_0 and η_0. In practice, the true coefficient structures are not available and thus making these two methods infeasible benchmarks. The simulation consists of 1000 repetitions and new x's, ζ's and υ's are drawn for each repetition,

Table 1: Monte Carlo simulation regression results

Estimator	**Mean Bias**	**Std. Dev.**	**rp(0.05)**	**# Selected variables**
Lasso	-0.694	0.084	1.000	n/a
Post-Lasso	-0.638	0.140	0.000	4.342
Indirect-Post-Lasso	-0.097	0.197	0.000	5.472
Post-Double-Lasso	-0.030	0.111	0.000	9.667
Double selection Oracle	0.002	0.209	0.004	n/a
Oracle	0.003	0.052	0.000	n/a

Note: Results are based on 1000 simulation repetitions with $p = 200$ and $n = 100$. The λ values for the Lasso variable selection steps are set according to the X-dependent rule as stated in Equation (8). The table presents mean bias over all simulation repetitions, the standard deviation of the treatment effect estimates from all simulation repetitions and rejection rates for 5% significance level tests. The rejection rates are computed using jack-knife standard errors to ensure robustness against heteroscedasticity (MacKinnon & White, 1985). The last column represents, if relevant, the average number of variables included in the final regression step. Implementation details of the estimators can be found in Appendix A.2

Table 1 displays mean bias, standard deviation, and rejection probabilities of 95% confidence intervals and thereby summarizes the inference performance of the studied estimation strategies.

Additionally, the last column displays the average number of selected variables that are included in the final regression step. In line with Belloni et al. (2011), both Lasso and post-Lasso show considerable mean bias. Indirect-post-Lasso, while showing minor bias relative to variability, demonstrates substantially higher variability than post-double-Lasso. Moreover, post-double-Lasso achieves the lowest mean bias amongst the feasible models while also maintaining relatively low variability among all 1000 estimates. The superior performance of post-double-Lasso could be expected as Lasso addresses regularization bias, while post-Lasso and indirect-post-Lasso suffer from substantial omitted variable bias due to the variable selection step being applied only to Equation (6) and Equation (7) respectively (Belloni et al., 2011). Belloni et al. (2014b) generalizes this result of post-double-selections superior performance by employing a diversity of simulation designs. A comparison of this thesis' results and those of Belloni et al. (2011), which conducted the same Monte Carlo simulation example, can be found in Appendix A.3.

The right column of Table 1 shows that the Lasso variable selection steps, using λ_X, impose strict penalties, selecting approximately 5% of the initial 200 predictors. Consequently, this thesis extends the simulation example by investigating two common λ determination methods that are less stringent to assess their effect on the performance of the estimation strategies. First, λ is determined through cross-validation and selected as the largest λ with the cross-validated error remaining within one standard error of the minimum error, referred to as λ_{se} from now. λ_{se} prioritizes model simplicity and interpretability over maximizing predictive accuracy, applying a more stringent penalty than the traditional cross-validation λ that minimizes the cross-validation error. Second, this thesis also employs the more traditional cross-validation λ, λ_{min}, which minimizes the cross-validation error. λ_{min} penalizes less stringent compared to λ_X and λ_{se}, which can be advantageous when the goal is to maximize predictive accuracy and simplicity and interpretability are less critical.

From the right column of Table 2 can be observed that the cross-validation techniques are less stringent, selecting approximately 9% and 22% of the initial 200 predictors. It is observed that, as more variables are included in the final regression, the performance of Lasso, post-Lasso, and indirect-post-Lasso improves, whereas the performance of post-double-Lasso deteriorates. This might be explained by the proportion of variables to observations in the final regression of post-double-Lasso, with the case λ_{min} involving on average 44 variables compared to only 100 observations. To conclude, the results indicate that stringent penalization in post-double-Lasso contributes substantially to its superior performance compared to the other considered estimation strategies. In line with this finding, Belloni et al. (2014b) note that the commonly used cross-validation λ_{min} might not ensure optimal performance in contexts where prediction is not the primary objective. Nevertheless, even when using cross-validation techniques to determine λ, post-double-Lasso maintains a significantly lower mean bias than Lasso and post-Lasso. Indirect-post-Lasso outperforms post-double-Lasso in terms of mean bias but still exhibits greater variability across their estimates compared to post-double-Lasso.

3.2 Double/debiased machine learning

Over time, ML has been recognized as an effective tool for predicting outcomes in complex, nonlinear environments and managing high-dimensional data (e.g. (Hastie et al., 2009)).

Table 2: Monte Carlo simulation regression results

Estimator	**Mean Bias**	**Std. Dev.**	**rp(0.05)**	**# Selected variables**
		λ_{se}		
Lasso	-0.551	0.121	0.983	n/a
Post-Lasso	-0.278	0.193	0.000	12.67
Indirect-Post -Lasso	-0.054	0.210	0.000	8.34
Post-Double-Lasso	-0.078	0.128	0.000	18.68
		λ_{min}		
Lasso	-0.397	0.132	0.635	n/a
Post-Lasso	-0.199	0.144	0.000	28.63
Indirect-Post-Lasso	0.023	0.267	0.000	21.48
Post-Double-Lasso	-0.097	0.232	0.000	44.24

Note: Results are based on 1000 simulation repetitions with $p = 200$ and $n = 100$. The λ values for the Lasso variable selection steps are set through cross-validation and selected as the largest λ with the cross-validated error remaining within one standard error of the minimum error (λ_{se}) and to minimize the cross-validation error (λ_{min}). The table presents mean bias over all simulation repetitions, the standard deviation of the treatment effect estimates from all simulation repetitions and rejection rates for 5% significance level tests. The rejection rates are computed using jack-knife standard errors to ensure robustness against heteroscedasticity (MacKinnon & White, 1985). The last column represents, if relevant, the average number of variables included in the final regression step. Implementation details can be found in Appendix A.2.

Nonetheless, ML estimators are prone to heavy regularization bias (Chernozhukov et al., 2018), therefore employing ML methods directly does not yield parameter estimates reflecting causal effects (Athey et al., 2018). Despite this, ML can be used for causal inference by segmenting the estimation process into several prediction tasks (Mullainathan & Spiess, 2017).

Chernozhukov et al. (2018) propose the double/debiased machine learning (DML) consisting of two stages. First, ML techniques are employed to estimate a set of so-called "nuisance functions". Second, utilizing the machine learning estimates of these nuisance parameters, the estimating equation is then solved to derive the target parameter estimates. DML generalizes the post-double-Lasso approach and eliminates the effects of overfitting and regularization bias by incorporating two fundamental elements: The first key element, Neyman orthogonal moments, is strongly related to the Frisch-Waugh-Lovell (FWL) theorem, a fundamental result in econometrics that provides a method for partialling out the effect of other variables while examining the effect of one specific variable in a linear regression model. Orthogonalization is now obtained by directly partialling out the effect of x from d in Equation (2), i.e. predicting treatment d from potential confounders x in Equation (2) and subsequently predicting y using the same variables in Equation (1). Essentially, an estimating equation satisfies Neyman orthogonality if it remains stable around the true nuisance functions, indicating that its average value is insensitive to slight variations in the nuisance functions (Shi et al., 2023). This insensitivity is characterized by a zero "derivative" of the expected value of the estimating equation with respect to the nuisance functions. Without the DML framework, the estimating equation for the true target variable α_0 can be described as follows:

$$E[\phi(W; \alpha_0, g_0(x_i))] = 0 \tag{9}$$

$$\text{where } \phi(W; \alpha_0, g_0(x_i)) = (y_i - d_i\alpha_0 - g(x_i))$$

where W is the data vector containing any coefficient value α and $g(x_i)$. The estimation equation of the DML procedure, enhances robustness against bias induced by ML nuisance estimation compared to Equation (9) in the following manner:

$$E[\phi(W; \alpha_0, l(x_i), f(x_i))] = 0 \tag{10}$$

$$\text{where } \phi(W; \alpha_0, l(x_i), f(x_i)) = [y_i - l(x_i) - \alpha_0(d_i - f(x_i))](d_i - f(x_i))$$
$$\text{and } l(x_i) = E(y_i|x_i), f(x_i) = E(d_i|x_i)$$

with $l(x_i)$ and $f(x_i)$ two new nuisance functions, not of direct interest, but readily estimable through ML methods. It can be shown that Equation (10) is Neyman orthogonal making it stable against small perturbations in nuisance functions and resilient to ML-induced bias (Shi et al., 2023). In other words, utilizing Neyman orthogonal moments removes the effect of regularization bias and diminishes sensitivity with respect to small pertuations in the nuisance parameters to estimate the coefficients (Chernozhukov et al., 2018).

The second key element involves cross-fitting, a sample-splitting procedure where the main and auxiliary samples are exchanged to produce multiple estimates, which are then averaged. This procedure reduces the probability of overfitting, an issue to which ML methods are prone when the same data is used for both training the model and evaluating its performance (Chernozhukov et al., 2018). Lastly, to enhance robustness concerning the random partitioning in finite samples, the process is reiterated for various splits, and the median estimate is reported. By employing these two key elements, a consistent and asymptotically normally distributed double ML estimator is obtained (Chernozhukov et al., 2018). To provide further clarity on the DML procedure, I outline Algorithm 1.

Algorithm 1 DML algorithm

1: Split the data into K folds
2: **for** $k = 1$ to K **do**
3: Train two machine learning models on $K - 1$ folds:
4: a) Outcome variable d (Equation (2))
5: b) Outcome variable y (Equation (1))
6: Employ these models to make predictions ($\hat{d}$ and $\hat{y}$) on the excluded fold.
7: Calculate residuals as $\hat{v} = d - \hat{d}$ and $\hat{\zeta} = y - \hat{y}$
8: Regress $\hat{\zeta}$ on $\hat{v}$, obtain the coefficient on $\hat{v}$
9: **end for**
10: Perform this process for each fold and average the coefficients to determine the final causal estimate.
11: Repeat the algorithm S times with varying splits (to ensure robustness concerning random partitioning in finite samples), then report the median estimate.

Recently, Fuhr et al. (2024) find that the suitability of specific ML methods for DML is strongly influenced by both the functional form of the confounding variables and the number

of confounders. They therefore warn against the use of Lasso, the most commonly applied ML method in DML, as it may yield biased estimates in the presence of nonlinear confounding without manual variable transformations. Additionally, the performance of DML Lasso deteriorates as the set of potential controls increases to handle these nonlinearities (Fuhr et al., 2024). Furthermore, their results suggest that DML with a flexible ML method is particularly effective at addressing nonlinear confounding without requiring knowledge of underlying functional forms, as opposed to simultaneously controlling for a large number of potentially important confounders.

Consequently, this thesis explores beyond Lasso, employing extreme gradient boosting (XGBoost) for the ML tasks within DML. XGBoost, introduced by Chen and Guestrin (2016), is an ensemble machine learning algorithm utilizing decision trees and gradient boosting and is a variant of the gradient boosting machine (GBM) algorithm. Ensemble machine learners combine multiple learners, called weak learners, to create a robust predictive model, aggregating their predictions. XGBoost utilizes decision trees as its base learners and forms an ensemble through boosting, sequentially adding weak learners to the ensemble with each new learner aiming to correct the errors made by the existing ones. The main difference from other GBM algorithms is the objective function, which is composed of two parts: The first term evaluates the model's predictive accuracy on the training data, known as the loss function. The second term, the so-called regularization term, controls the complexity of the model, which improves interpretability. The objective function can be represented as follows:

$$\mathcal{L} = \sum_i \iota(\hat{y}_i, y_i) + \sum \tau(f) \tag{11}$$

$$\text{where } \tau(f) = \gamma T + \frac{1}{2}\lambda ||w||^2$$

with γ controlling the penalty for the numbers of leaves T, and λ controlling the penalty for the magnitude of leaf weights w, for each tree f. The loss function, $\sum_i \iota(\hat{y}_i, y_i)$, utilizes Newton boosting instead of gradient boosting by applying a second-order Taylor expansion, which is a key factor contributing to its enhanced performance (Aras & Lisboa, 2022). For an in-depth explanation of XGBoost, this thesis refers to Chen and Guestrin (2016). The performance of XGBoost has attained widespread recognition across numerous machine learning and data mining challenges (Chen & Guestrin, 2016). More specifically, Shi et al. (2023) recognize XGBoost as a powerful method to use within the DML framework. Moreover, Fuhr et al. (2024) find DML using XGBoost performing excellent across diverse settings, even leading them to advocate for its adoption as the default method within DML. Their main reason for this is that XGBoost DML performs superior across a diverse set of simulation, including one simulation where the functional form of each variable is randomly drawn, closely resembling real-world settings.

3.2.1 Monte Carlo Simulation

To assess performance relative to the earlier described and employed models in the simulation example (Section 3.1.1), DML Lasso and DML XGBoost are applied to simulated data under similar conditions. Due to computational time constraints, the number of simulation repetitions is set to 100 instead of 1000, while all other settings remain unchanged.

Table 3 demonstrates that the same pattern observed in Table 1 persists for the first six models, indicating that this pattern is robust to the adjustment in the number of simulation repetitions. DML Lasso achieves the lowest mean bias, closely approaching the results of the infeasible Oracle models, while maintaining relatively low variance. The superior performance of DML Lasso could be expected as Fuhr et al. (2024) concluded on DML Lasso performing well in a similar simulation setting, with the true relations being linear. In contrast, DML XGBoost only outperfoms Lasso and post-Lasso and achieves similar performance to indirect-post-Lasso. Consequently, DML XGBoost is outperformed by both post-double-Lasso and DML Lasso. This can be explained by the true structure of the coefficients, β and η, and how XGBoost handles this structure. Tree-based methods perform well in modelling step functions and interactions but need more data to effectively model linear functions (James, Witten, Hastie, Tibshirani & Taylor, 2023). Corresponding, Fuhr et al. (2024) employ multiple simulations and find that flexible DML methods, such as XGBoost, exhibit increased bias in small samples with linear confounding but rapidly understand the linear relationships as the sample size increases. Contrary to its poor performance in small samples with only linear functional forms, DML XGBoost demonstrates superior performance in simulations that more closely resemble real-world settings, where the functional forms of variables are drawn randomly (Fuhr et al., 2024). In conclusion, the pattern in model performance showed in Table 3 is strongly dependent on the true functional forms in the simulation and does not necessarily reflect the performance in real world settings.

Table 3: Monte Carlo simulation regression results

Estimator	**Mean Bias**	**Std. Dev.**
Lasso	-0.690	0.085
Post-Lasso	-0.648	0.132
Indirect-Post-Lasso	-0.085	0.195
Post-Double-Lasso	-0.033	0.113
Double Oracle	0.003	0.113
Oracle	0.005	0.104
DML Lasso	-0.009	0.125
DML XGBoost	-0.100	0.120

Note: Results are based on 100 simulation repetitions with $p = 200$ and $n = 100$. The table presents mean bias over all simulation repetitions and the standard deviation of the treatment effect estimates from all simulation repetitions. Implementation details on the estimators can be found in Appendix A.2.

3.3 Method implementation

Finally, this thesis employs and compares the following models when studying the abortion and crime hypothesis: Post-double-Lasso, DML Lasso and DML XGBoost. Lasso is implemented using the *glmnet* R-package (Simon, Friedman, Hastie & Tibshirani, 2011). For post-double-Lasso, the regularization parameter λ is determined by 10-fold cross-validation and selected as λ_{se}, which is the largest λ with the cross-validated error remaining within one standard error of the minimum. By using a more stringent penalty, λ_{se} favors model simplicity and interpretability over the maximum predictive accuracy achieved by the conventional cross-validation λ_{min}, which minimizes the cross-validation error and penalizes less stringent. For DML Lasso, λ is determined

through a grid search from 0.0001 to 1 using 5-fold cross-validation. The grid search is chosen over the more conventional automatic internal cross-validation is this case due to time limitations, as the DML algorithm already involves multiple folds and repetitions. XGBoost is implemented using the *xgboost* R-package of Chen et al. (2024). In line with (Fuhr et al., 2024), the number of boosting iterations is determined through a grid search using 5-fold cross-validation, exploring up to 200 rounds. In addition, the maximum depth of the individual trees in the ensemble models and the learning rate are tuned using 5-fold cross-validation through a grid search around their default values. Finally, the implementation of both DML models, is done following Bach, Chernozhukov, Kurz, Spindler and Klaassen (2021), a general implementation of the DML approach of Chernozhukov et al. (2018). The number of folds K is set to five and the algorithm gets repeated 100 times ($S = 100$), as recommended by Chernozhukov et al. (2018) when applying DML in real world applications. More detailed implementation information can be found in my R-code, which is provided in the supplementary material.

4 Results

Table 4 represents three different specifications for each of the three crime types, reporting the treatment effect estimate, its significance, and its standard error. Alongside the models and results, I include the findings of Donohue III and Levitt (2001) and (Donohue & Levitt, 2020) for comparison.

Similar to Donohue and Levitt (2020), all coefficients on abortion for violent- and property crime rates are negative for each time period, indicating that higher abortion rates correspond to declining violent- and property crime rates. The significance and magnitude of the estimates differ over time. When focusing on violent crime, two out of three of my models find significant negative results for the first sub-sample period and the full sample period. In contrast, for the second sub-sample period, only one of the models find a significant result. Together with the decreased magnitude, this could indicate that the effect of abortion on violent crime rates has weakened over time and does not hold significantly when only considering the time period from 1998-2014. However, the significance found by DML XGBoost and post-double-Lasso in the full sample period indicates that the original period's significance holds true for the entire sample period. When focusing on property crime, most of the models identify significant negative effects. In the second sub-sample period, all models show significant results. The increased magnitude observed by post-double-Lasso and DML XGBoost suggests that the original period's effect strengthens and persists in the second sub-sample period. Interestingly, only one model yields a significant result for the full sample period, and all three models demonstrate a decreased magnitude. Across all time periods and models, no significant result for the effect of abortion on murder crime rates is obtained. In the second sub-sample period, post-double-Lasso and DML XGBoost even find small positive effects instead of the common discussed negative effect. In line with the previous, the reported standard errors are substantially higher compared to the other two crime types. A possible explanation could be related to the UCR data on murder crime rates, which exhibit significant undercounts, especially in states with low abortion rates (Donohue & Levitt, 2020). Nonetheless, this does not account for the difference in significance between Donohue III and Levitt (2001), Donohue and Levitt (2020), and this thesis' results, as

Table 4: Estimates of the effect of abortion- on crime rates

Crime	**Violent**		**Property**		**Murder**	
Model	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**
1985 - 1997						
Donohue III and Levitt (2001)	-0.129°	0.024	-0.091°	0.018	-0.121°	0.047
Donohue and Levitt (2020)	-0.178**	0.022	-0.152**	0.016	-0.100*	0.040
Post-Double-Lasso	-0.165***	0.037	-0.079***	0.018	-0.165	0.148
DML Lasso	-0.319***	0.044	-0.118***	0.026	-0.296	0.072
DML XGBoost	-0.071	0.037	-0.042	0.023	-0.135	0.115
1998 - 2014						
Donohue and Levitt (2020)	-0.189**	0.019	-0.168**	0.015	-0.152**	0.021
Post-Double-Lasso	-0.138***	0.048	-0.119***	0.042	0.001	0.123
DML Lasso	-0.089	0.033	-0.092**	0.033	-0.027	0.078
DML XGBoost	-0.049	0.047	-0.115*	0.053	0.031	0.131
1985 - 2014						
Donohue and Levitt (2020)	-0.189**	0.019	-0.165**	0.015	-0.158**	0.020
Post-Double-Lasso	-0.146***	0.030	-0.055***	0.017	-0.093	0.092
DML Lasso	-0.054	0.076	-0.030	0.017	-0.038	0.073
DML XGBoost	-0.094***	0.030	-0.033	0.023	-0.083	0.079

Note: This table reports results from estimating the effect of abortion rate on three different crime rates. The table is divided into three multi-columns, each representing a different dependent variable. Within each multi-column, column 'Estimate' represents the estimate of the effect of abortion rate on the relevant dependent variable, and the column 'SE' represents the standard error of this coefficient. Donohue III and Levitt (2001) and Donohue and Levitt (2020) derive their standard errors using the sandwich estimator proposed by White (1980), which adjusts for heteroskedasticity and possible correlation in the error terms. This thesis derives the standard errors using the asymptotic standard error, $\hat{\sigma}/\sqrt{n}$. *, ** and *** indicate significance at the 5%, 1% and 0.5% level respectively. As Donohue III and Levitt (2001) do not indicate at which level their estimates are 'highly significant', their estimates are presented with ∘. Additionally, the table rows are divided into three parts, each one representing a time period.

they still obtain significant results when using UCR murder crime rate data.

To assess the effectiveness of my three models in estimating the treatment effect, this and the following two paragraphs compare results across models. The models obtain different results in terms of magnitude and significance. During the first sub-sample period, DML Lasso obtains estimates notably larger than post-double-Lasso and DML XGBoost. In addition, in the second sub-sample period, DML Lasso uniquely shows a decrease in magnitude from the first period for the property crime rate specification and is the only model to yield a negative estimate for the murder crime type specification. Remarkably, DML XGBoost yields no significant estimates during the first sub-sample period while the other two models do for both violent- and property crime. In contrast, post-double-Lasso frequently yields significant estimates, namely six out of nine specifications, while DML Lasso and DML XGBoost obtain significant estimates only three and two times respectively. Nonetheless, there are also similarities across models, as they all agree on a significant negative treatment effect for property crime in the second sub-sample period. Furthermore, some models yield approximately similar estimates in certain specifications (e.g., post-double-Lasso and DML XGBoost for property crime in the second sub-sample period).

Determining which model captures the true treatment effect the best is challenging, as a fundamental complication in causal inference is the uncertainty surrounding the unknown true effect in real world applications. Fuhr et al. (2024) investigate whether the predictive accuracy of different ML methods in predicting the treatment- and outcome variable aligns with their capability of recovering the true treatment effect within the DML approach. This hypothesis assumes that methods effectively modeling confounding relationships would reduce bias in effect estimates compared to less accurate estimators. Under the strong assumption that the methods account for all relevant confounders, the predictive accuracy of ML methods can be used as a criterion for determining the reliability of each estimate (Fuhr et al., 2024). In alignment with Fuhr et al. (2024), Shi et al. (2023) observe a trend indicating that increased ML predictive errors are associated with less accurate estimates of the treatment effect. Therefore, Table 5 displays the mean squared error (MSE) of the methods when predicting the treatment d in Equation (2) and outcome y in Equation (1). The bold values in the table indicate the lowest MSE value for each specification across the three different crime types and sample periods.

Table 5: Mean squared errors of the prediction steps within the post-double-selection and DML approach

Crime	**Violent**		**Property**		**Murder**	
Model	**MSE(Y)**	**MSE(D)**	**MSE(Y)**	**MSE(D)**	**MSE(Y)**	**MSE(D)**
1985 - 1997						
Post-Double-Lasso	**0.006**	**0.007**	0.003	0.015	**0.063**	**0.005**
DML Lasso	0.009	0.011	0.003	0.016	0.064	0.010
DML XGBoost	**0.006**	**0.007**	**0.002**	**0.012**	0.070	**0.005**
1998 - 2014						
Post-Double-Lasso	**0.005**	**0.003**	**0.002**	**0.001**	**0.042**	0.004
DML Lasso	**0.005**	0.005	**0.002**	0.002	0.043	0.005
DML XGBoost	**0.005**	**0.003**	**0.002**	**0.001**	0.045	**0.003**
1985 - 2014						
Post-Double-Lasso	**0.006**	**0.005**	**0.002**	0.007	**0.053**	**0.004**
DML Lasso	0.007	0.007	0.003	0.011	**0.053**	0.006
DML XGBoost	**0.006**	**0.005**	**0.002**	**0.006**	0.058	0.005

Note: This table reports the MSE of the prediction steps within the post-double-selection and DML approached. The table is divided into three multi-columns, each representing a different dependent variable. Within each multi-column, column MSE(Y) represents the MSE when predicting the outcome y in Equation (1) and MSE(D) represents the MSE when predicting d in Equation (2). The bold values indicate the lowest MSE value for the specification in question.

When analyzed by period, DML XGBoost emerges as the most reliable model for the first sub-sample period, achieving the lowest MSE in five out of six specifications. This finding indicates that the relationships between the variables and abortion- and crime rates during the first sub-sample period are likely nonlinear and complex, which can be effectively modeled by XGBoost and less effectively by Lasso. In the second sub-sample period, both post-double-Lasso and DML XGBoost show comparable reliability, achieving the lowest MSE in five out of six specifications. However, in the full sample period, post-double-Lasso consistently achieves the lowest MSE except for one specification where DML XGBoost achieves the lowest MSE. When analyzed by

crime type, for property crime, DML XGBoost achieves the lowest MSE for each specification against three and one time for post-double-Lasso and DML Lasso respectively. Indicating possible nonlinear and highly complex relations between the variables and the treatment- and dependent variable, which is where XGBoost's performance benefits from compared to Lasso. When focusing on crimes of type murder, post-double-Lasso outperforms the other models, achieving the lowest MSE in all specifications except one. This finding suggest that for crime type murder, the relations across variables are likely to be more linear. Overall, post-double-Lasso and DML XGBoost perform similarly well, outperforming each other in some contexts but mostly performing similar. What stands out is the bad performance of DML Lasso, achieving only three times the lowest MSE across all eighteen specifications. The performance difference of Lasso between post-double-Lasso and DML Lasso likely arises from their tuning processes. Post-double-Lasso does not use sample splitting or algorithm repetitions, while DML Lasso involves five fold sample splitting and 100 repetitions of the whole algorithm, allowing less time for tuning Lasso for its prediction steps.

To contextualize my findings, the results are now compared to the findings of Donohue III and Levitt (2001) and Donohue and Levitt (2020) using Table 4. When focusing at the first sub-sample period, all models agree on the treatment effect being negative. However, not all of the models agree on the high significance of the estimates found by Donohue III and Levitt (2001) and Donohue and Levitt (2020). Moreover, the effects' magnitude differs across models, with DML Lasso yielding particularly notable results as they are outstandingly large for violent- and murder crime rates. Nevertheless, the MSE analysis from Table 5 shows that DML Lasso is less reliable in these specifications thereby favoring reliability of the results obtained by post-double-Lasso and DML XGBoost. In contrast to Donohue and Levitt (2020), which obtain larger estimates for the second sub-sample period compared to the first sub-sample period, the findings of this thesis do not indicate larger coefficients for the second time period across all three crime types. All three models obtain estimated effects substantially lower, in absolute terms, than those obtained in the first time period for the violent- and murder crime rate specifications. Contrary to the previous but in line with the over time increasing trend in Donohue and Levitt (2020), for the property crime specifications, post-double-Lasso and DML XGBoost achieve estimates approximately twice as high compared to the first sub-sample time period. For the full time period and across both violent and property crime rate specifications, this thesis confirms the significant negative effect of abortion on crime rates, consistent with the findings of Donohue and Levitt (2020). Despite this, the estimated coefficients of these effects are substantially smaller then those reported by Donohue and Levitt (2020). The biggest differences between this thesis and the studies conducted by Donohue III and Levitt (2001) and Donohue and Levitt (2020), lie in the findings regarding the murder crime rates. In contrast to both studies, which report a significant effect during the first sub-sample period, none of my models yield a significant effect, with all models producing notably large standard errors. Additionally, over the second sub-sample period, both post-double-Lasso and DML XGBoost find a positive treatment effect which is in contrast to the substantially bigger negative significant effect found by Donohue and Levitt (2020). Consistent with the two sub-sample periods but in contrast with Donohue and Levitt (2020), my models do not obtain significant results for the murder crime rate specification

for the full time period.

Belloni et al. (2014b) and Belloni et al. (2014a) performed a similar replication study using post-double-Lasso, data from 1985-1997, and similarly accounting for higher-order and interaction terms. Although there are many similarities in methodology and data, this thesis produces different results when considering the standard errors and significance. Table 6 reports the results of this thesis and those of Belloni et al. (2014b) and Belloni et al. (2014a). While both studies found no significant results using post-double-Lasso for any of the three crime types, this thesis finds significant results across both violent- and property crime rates and obtains standard errors about three times smaller than the ones reported by Belloni et al. (2014b) and Belloni et al. (2014a). In contrast, for the murder crime rate specifications, my results do align with those of both studies. A possible explanation for the differences between this thesis and the studies conducted by Belloni et al. (2014b) and Belloni et al. (2014a) could be the level of penalization within the two variable selection steps. Belloni et al. (2014b) selects eight, twelve, and nine variables out of 284 potential controls, whereas this thesis selects 34, 13 and 20 variables out of 150 potential controls, for violent-, property-, and murder crime rates respectively. The level of penalization strongly depends on how λ in Equation (5) is determined, which is where my approach differs from Belloni et al. (2014b) and Belloni et al. (2014a). Moreover, the differences might stem from the way standard errors are computed. Whereas both Belloni et al. (2014b) and Belloni et al. (2014a) compute clustered standard errors at the state level, this thesis' standard errors do not account for the within-state correlation.

Table 6: Estimates of the effect of abortion- on crime rates, comparing post-double-Lasso results

Crime	**Violent**		**Property**		**Murder**	
Model	**Estimate**	**SE**	**Estimate**	**SE**	**Estimate**	**SE**
1985 - 1997						
Belloni et al. (2014b)	-0.104	0.107	-0.030	0.055	-0.125	0.151
Belloni et al. (2014a)	-0.171	0.117	-0.061	0.057	-0.189	0.177
Post-Double-Lasso	-0.165***	0.037	-0.079***	0.018	-0.165	0.148

Note: This table reports results from estimating the effect of abortion rate on three different crime rates. The table is divided into three multi-columns, each representing a different dependent variable. Within each multi-column, column 'Estimate' represents the estimate of the effect of abortion rate on the relevant dependent variable and the column 'SE' represents the standard error of this coefficient. Belloni et al. (2014b) and Belloni et al. (2014a) compute their standard errors clustered at the state level. This thesis derives the standard errors using the asymptotic standard error, $\hat{\sigma}/\sqrt{n}$. *** indicates significance at the 0.5% level.

5 Conclusion

This thesis studies the causal effect of abortion legalization on crime rates in the U.S. Since the original conducted study by Donohue III and Levitt (2001), this has become a widely studied causal inference issue where researchers have not yet reached a consensus, as conclusions vary across different studies. The effect of abortion legalization on crime rates is examined over a full sample period from 1985 to 2014 and two sub-sample periods: 1985 to 1997 and 1998 to 2014. Furthermore, three dependent variables are considered, each one representing another type of crime: violent-,

property- and murder crime. Additionally, this thesis compares multiple models and their causal inference through a simulation example. Specifically, it employs the post-double-selection method, introduced by Belloni et al. (2014b), and the double debiased machine learning (DML) method, introduced by Chernozhukov et al. (2018). The post-double-selection methods is employed using Lasso, resulting in the post-double-Lasso method. The DML approach is applied using two distinct ML methods: Lasso and XGBoost.

My results present negative estimates for the causal effect, indicating that higher abortion rates indeed correspond to declining crime rates. However, not all models yield significant results in every specification. Post-double-Lasso and DML XGBoost strongly outperform DML Lasso in predictive accuracy, making their estimates more reliable. The choice between post-double-selection and DML XGBoost varies by time period and crime type, as they each excel in different specifications but generally show comparable performance. The primary difference between the results of those two models lies in the significance of their results. Whereas post-double-Lasso obtains significant estimates in six out of nine specifications, DML XGBoost yields significant estimates in only two cases. Additionally, compared to post-double-Lasso, DML XGBoost consistently produces estimates of smaller magnitude across all specifications.

For murder crime rates, none of my models obtains significant results for any of the time periods, indicating that I cannot conclude on a significant negative effect of higher abortion rates on murder rates. Post-double-Lasso and DML XGBoost even obtain positive effect estimates for the period ranging from 1998 to 2014, making the hypothesis of Donohue III and Levitt (2001) for crime type murder even more questionable. For the violent- and property crime specifications, significant effects are obtained for each time period by at least one of my methods. In conclusion, this thesis aligns with previous findings regarding the significant impact of abortion legalization on violent and property crime rates. However, the observed effects are of a smaller magnitude compared to those reported in the original studies by Donohue III and Levitt (2001) and Donohue and Levitt (2020), and not all methods agree on the significance of these effects. Finally, finding no evidence for a significant effect of abortion rate on murder crime rate, this thesis' results do not align with those of Donohue III and Levitt (2001) and Donohue and Levitt (2020) with respect to the murder crime rate specification.

This thesis compares models based on predictive accuracy, operating under the strong assumption that all relevant confounders are accounted for and that modeling these confounders effectively reduces bias in eventual effect estimates. However, in real-world applications, this is a strong assumption, and this hypothesis is at odds with growing literature indicating that obtaining high predictive accuracy differs fundamentally from the aim of causal inference. A suggestion for further research is therefore to employ a simulation example that closely resembles the relevant real world application. For instance, a simulation example in which the functional forms are randomly drawn out of a diverse set of simple and complex forms, and where the number of variables and observations is based on the actual dataset. Using this simulation, investigating the use of two distinct ML methods within a DML approach could be beneficial for the treatment effect estimate; as one method might excel in the predicting outcome while the other may be better suited for the predicting treatment, dependent on the functional forms of the variables. Secondly, to further investigate the predictions of Donohue III and Levitt (2001)

and Donohue and Levitt (2020) of the advocated effect in the future, it could be valuable to extend the dataset by incorporating data from more recent years.

Finally, the tuning processes for the methods examined in this thesis could be optimized. For instance, this thesis repeats the DML algorithm 100 times, in line with the recommendation of Chernozhukov et al. (2018) for real-world contexts. However, Fuhr et al. (2024) performed a stability analysis for their real-world application and found the necessity of repeating the algorithm 199 times to achieve stable results. For the abortion and crime hypothesis, it could be beneficial to perform a similar analysis and determine at which number of repetitions the DML methods obtain stable results. In addition, due to time constraints, the tuning of the ML method itself within the DML approach is now performed for a limited number of hyperparameters using a grid search through cross-validation. For further research, I recommend a more thorough and detailed tuning procedure. Lastly, my results using post-double-Lasso for the abortion and crime hypothesis differ from those of Belloni et al. (2014b) and Belloni et al. (2014a) using the same method and similar data. The simulation employed in this thesis indicates that post-double-Lasso's performance varies substantially based on the level of penalization. Therefore, simulating a scenario resembling this particular application could help determine the optimal tuning of the penalization term.

References

Aras, S. & Lisboa, P. J. (2022). Explainable inflation forecasts by machine learning models. *Expert systems with applications*, *207*, 117982.

Athey, S. et al. (2018). The impact of machine learning on economics. *The economics of artificial intelligence: An agenda*, 507–547.

Bach, P., Chernozhukov, V., Kurz, M. S., Spindler, M. & Klaassen, S. (2021). Doubleml–an object-oriented implementation of double machine learning in r. *arXiv preprint arXiv:2103.09603*.

Belloni, A. & Chernozhukov, V. (2011). High dimensional sparse econometric models: An introduction. *arXiv e-prints*, arXiv–1106.

Belloni, A. & Chernozhukov, V. (2013). Least squares after model selection in high-dimensional sparse models. *Bernoulli*, *19*(2), 521–547.

Belloni, A., Chernozhukov, V. & Hansen, C. (2011). Inference for high-dimensional sparse econometric models. *arXiv preprint arXiv:1201.0220*.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014a). High-dimensional methods and inference on structural and treatment effects. *Journal of Economic Perspectives*, *28*(2), 29–50.

Belloni, A., Chernozhukov, V. & Hansen, C. (2014b). Inference on treatment effects after selection among high-dimensional controls. *Review of Economic Studies*, *81*(2), 608–650.

Bijmolt, T. H., Van Heerde, H. J. & Pieters, R. G. (2005). New empirical generalizations on the determinants of price elasticity. *Journal of marketing research*, *42*(2), 141–156.

Bilancini, E., Boncinelli, L., Di Paolo, R., Menicagli, D., Pizziol, V., Ricciardi, E. & Serti, F. (2022). Prosocial behavior in emergencies: Evidence from blood donors recruitment and retention during the covid-19 pandemic. *Social Science & Medicine*, *314*, 115438.

Bryan, C. J., Yeager, D. S. & O'Brien, J. M. (2019). Replicator degrees of freedom allow publication of misleading failures to replicate. *Proceedings of the National Academy of Sciences*, *116*(51), 25535–25545.

Card, D. (1999). The causal effect of education on earnings. *Handbook of labor economics*, *3*, 1801–1863.

Chen, T. & Guestrin, C. (2016). Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining* (pp. 785–794).

Chen, T., He, T., Benesty, M., Khotilovich, V., Tang, Y., Cho, H., ... others (2024). Xgboost: extreme gradient boosting. *R package version 1.7.7.1*, *1*(4), 1–4.

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

Comanor, W. & Phillips, L. (1995). The impact of income and family structure on delinquency.

Donohue, J. J. & Levitt, S. (2020). The impact of legalized abortion on crime over the last two decades. *American law and economics review*, *22*(2), 241–302.

Donohue, J. J. & Levitt, S. D. (2004). Further evidence that legalized abortion lowered crime: A reply to joyce. *Journal of Human Resources*, *39*(1), 29–49.

Donohue III, J. J. & Levitt, S. D. (2001). The impact of legalized abortion on crime. *The Quarterly Journal of Economics*, *116*(2), 379–420.

Fuhr, J., Berens, P. & Papies, D. (2024). Estimating causal effects with double machine learning–a method evaluation. *arXiv preprint arXiv:2403.14385*.

Hastie, T., Tibshirani, R. & Friedman, J. H. (2009). *The elements of statistical learning: data mining, inference, and prediction* (Vol. 2). Springer.

James, G., Witten, D., Hastie, T., Tibshirani, R. & Taylor, J. (2023). *An introduction to statistical learning: With applications in python*. Springer Nature.

Joyce, T. (2006). Further tests of abortion and crime: A response to donohue and levitt (2001, 2004, 2006). *NBER Working Paper*(w12607).

Leeb, H. & Pötscher, B. M. (2008). Can one estimate the unconditional distribution of post-model-selection estimators? *Econometric Theory*, *24*(2), 338–376.

Levine, P. B., Staiger, D., Kane, T. J. & Zimmerman, D. J. (1999). Roe v wade and american fertility. *American Journal of Public Health*, *89*(2), 199–203.

MacKinnon, J. G. & White, H. (1985). Some heteroskedasticity-consistent covariance matrix estimators with improved finite sample properties. *Journal of econometrics*, *29*(3), 305–325.

Mullainathan, S. & Spiess, J. (2017). Machine learning: an applied econometric approach. *Journal of Economic Perspectives*, *31*(2), 87–106.

Shi, B., Mao, X., Yang, M. & Li, B. (2023). What, why, and how: An empiricist's guide to double/debiased machine learning. *Debiased Machine Learning (December 27, 2023)*.

Shmueli, G. (2010). To explain or to predict? *Statistical Science*, *25*(3), 289–310.

Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. (2011). Regularization paths for cox's proportional hazards model via coordinate descent. *Journal of Statistical Software*, *39*(5).

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, *58*(1), 267–288.

White, H. (1980). A heteroskedasticity-consistent covariance matrix estimator and a direct test for heteroskedasticity. *Econometrica: journal of the Econometric Society*, 817–838.

Woody, S., Carvalho, C. M., Hahn, P. R. & Murray, J. S. (2020a). Estimating heterogeneous effects of continuous exposures using bayesian tree ensembles: revisiting the impact of abortion rates on crime. *arXiv preprint arXiv:2007.09845*.

Woody, S., Carvalho, C. M. & Murray, J. S. (2020b). Bayesian inference for treatment effects under nested subsets of controls. *arXiv preprint arXiv:2001.07256*.

Zimring, W. D. S. F. E. et al. (2006). *The great american crime decline*. Oxford University Press, USA.

A Monte Carlo Simulation

A.1 Iteration procedure σ for X-dependent λ

In practice, the conservative estimate for σ, $\bar{\sigma} = \sqrt{\text{var}_n(y_i)} = \sqrt{\mathbb{E}_n[(y_i - \bar{y})^2]}$ where $\bar{y} = \mathbb{E}_n[y_i]$, often leads to overpenalization via Equation (8) (Belloni et al., 2011). Therefore, I utilize the iteration procedure, described in (Belloni & Chernozhukov, 2011), to estimate σ. This procedure can be described as followed: Define $\hat{\sigma}_{I_0} = \sqrt{E[(y_i - x_i'\bar{\beta}(I_0))^2]}$ where I take I_0 as the set

including only the intercept and $\bar{\beta}(I_0)$ the least squares estimator of the coefficients associated with I_0. Then the iterative procedure described in Algorithm 2 is employed and the resulting $\hat{\sigma}$ is used to estimate λ in Equation (8).

Algorithm 2 Iteration procedure to estimate σ

1: Choose a positive number ψ, set $\hat{\sigma}_0 = \psi\hat{\sigma}_{I_0}$ and specify a small constant $\nu \geq 0$. Set $k = 0$ and specify a constant $K > 1$ as an upperbound on the number of iterations.
2: **for** $k = 1$ to K **do**
3: Compute the Lasso estimator, where the starting λ_X is computed using $\hat{\sigma}_{I_0}$
4: $\hat{\sigma}^2_{k+1} = \hat{\sigma}(\hat{\beta})$
5: If $|\hat{\sigma}_{k+1} - \hat{\sigma}_k| \leq \nu$, a small constant ≥ 0 or $k > K$, report $\hat{\sigma} = \hat{\sigma}_{k+1}$; otherwise set $k \leftarrow k + 1$ and go to 3.
6: **end for**

A.2 Implementation in R

For the models Lasso, post-Lasso, indirect-post-Lasso and post-double-Lasso, the *glmnet* package is used (Simon et al., 2011). As this package employs $1/n$ parametrization (Simon et al., 2011), λ_X is scaled by n before being passed to the *glmnet* Lasso model. λ_{min} and λ_{se} are determined via the automatic internal cross-validation of *glmnet* and therefore do not require scaling. Both DML methods are implemented using the *DoubleML* R-package of Bach et al. (2021). Within the DoubleML R-models, Lasso is implemented using the *glmnet* R-package (Simon et al., 2011) and XGBoost is implemented using the *xgboost* R-package. Following the default settings and for simplicity purpose, the regularization parameter for Lasso is determined by cross-validation and selected as the λ that minimizes the cross-validation error. For XGBoost, the number of boosting iterations and the learning rate are tuned through five-fold cross validation using a grid search of 1 to 200 and 0.01 and 0.3 respectively. The grid search for the maximum value of the number of boosting iterations is determined based on the setting used in the research by Fuhr et al. (2024), while the learning rate values for the grid search are chosen around its default value. Second to last, following the recommendation of (Chernozhukov et al., 2018), the number of folds for the DML algorithm is set to five. Lastly, the DML algorithm gets repeated 10 times ($S = 10$). This number of repetitions is based on Fuhr et al. (2024), which conducted a simulation study under similar settings and observe stable estimates after approximately nine algorithm repetitions.

A.3 Comparison with Belloni et al. (2011)

Section 3.1.1 conducts the same Monte Carlo simulation as performed by (Belloni et al., 2011). Table 7 presents the simulation results of Belloni et al. (2011) and this thesis' simulation results (within brackets). Whereas this thesis shows a similar pattern in terms of mean bias and standard deviation across models, the actual values differ. Whereas similar mean bias and standard deviation values are found for Lasso, post-Lasso and indirect-post-Lasso, from post-double-Lasso onwards this thesis finds mean biases ten times higher than those reported by Belloni et al. (2011). In addition, there are notable differences observed when comparing rejection rates.

As all known settings in both simulation examples are identical, the discrepancies must stem from discretionary choices that I had to make during this replication process. For instance, the

iteration procedure for σ used for the computation of λ_X requires choices for parameters ψ, ν and K (see Algorithm 2). Furthermore, it is unclear how Belloni et al. (2011) exactly computes the rejection rates. I therefore choose to determine those by computing Z-scores for each simulation repetition by dividing the bias by the jack-knife standard error (MacKinnon & White, 1985) and count for how much repetitions this Z-scores are larger than 1.96 (because of the 5% confidence level). It is possible that Belloni et al. (2011) may employ another method to compute the standard errors, potentially leading to different rejection rates.

Table 7: Monte Carlo simulation regression results, a comparison

Estimator	**Mean Bias**		**Std. Dev.**		**rp(0.05)**	
Lasso	0.644	(-0.694)	0.093	(0.084)	1.000	(1.000)
Post-Lasso	0.415	(-0.638)	0.209	(0.140)	0.877	(0.000)
Indirect-Post-Lasso	0.091	(-0.097)	0.194	(0.197)	0.004	(0.000)
Post-Double-Lasso	-0.004	(-0.030)	0.111	(0.111)	0.054	(0.000)
Double selection Oracle	0.0001	(0.002)	0.110	(0.209)	0.051	(0.004)
Oracle	-0.0003	(0.003)	0.100	(0.052)	0.044	(0.000)

Note: Results are based on 1000 simulation repetitions with p = 200 and n = 100. The λ values for the Lasso variable selection steps are set according to the X-dependent rule as stated in Equation (8). The table presents mean bias over all simulation repetitions, the standard deviation of the treatment effect estimates from all simulation repetitions and rejection rates for 5% significance level tests. The rejection rates within brackets are formed using jack-knife standard errors which are robust to heteroscedasticity (MacKinnon & White, 1985). The columns without brackets represent results from Belloni et al. (2011), the values within brackets represent this thesis' results.