

ECO 395M: Final Project

Using Machine Learning literature to predict counterfactuals: an alternative method to Differences-in-Differences estimation

Bernardo Arreal Magalhaes - UTEID ba25727

Adhish Luitel - UTEID al49674

Ji Heon Shim - UTEID js93996

Abstract

In this project, we looked into the effect of prison expansion across various prisons in the state of Texas. To assess the causal effect of increased prison capacity on black male population incarceration, we first used the Differences-in-Differences estimator and argued about the validity of the parallel trends assumption. Then, in order to overcome potential weaknesses of the Diff-in-Diff model in this research project, we explored three alternative predictive analysis models, i.e. Lasso regression, Random Forests, and Boosting model respectively to test their performances via K-fold validation to determine the most accurate model. As a result, our model fits data quite accurately suggesting the high possibility of predicting counterfactuals.

Introduction

Assessing policy effects and making predictions based on it has always been a key part of quantitative economics. Empirical economists are very often interested in estimating the impact of certain events or policies on a particular outcome. Wooldridge (2013) describes the effectiveness applications of Differences-in-Differences methodology when the data arise from a natural experiment. This kind of experiment occurs when an exogenous event changes the environment in which individuals operate, and require observations of both treatment and control group before and after the change.

This methodology is particularly powerful for inferring causality since it neutralizes unobserved, but fixed, omitted variables (Angrist and Pischke, 2018). Nonetheless, it relies on a quite strong – and unfortunately not testable – assumption that the outcome in the multiple individuals/firms/states included in the analysis share the same trend over time, which is called parallel trends assumption.

The table below illustrates a simple version of the Diff-in-Diff method and why this assumption is required. By decreasing the outcome after treatment from the outcome before treatment for a treated state, the difference (D1) is going to be the effect caused by the treatment (E) plus a trend (T). This step neutralizes unobserved factors of a particular state. For a state that wasn't treated, the difference (D1) before and after treatment is the trend (T) only. So, if we assume that T is the same for both states, we can decrease T, that was measured from the control state, from $T + E$ in order to isolate the causal effect E.

Table 1: The basic concept of diff-in-diff

State	Time	Outcome	D1	D2
Treated State	Before	$Y = S_t$	$T + E$	E
	After	$Y = S_t + T + E$		
Control State	Before	$Y = S_c$	T	E
	After	$Y = S_c + T$		

This assumption is not testable because we don't know what would've happened to the treatment state had it never been treated.

But what if we could predict what would have happened to the treated state in this alternative world where it wasn't treated without having to rely on the parallel trends assumption?

In order to do that, we analyzed the application of a set of predictive models such as Lasso Regression, RandomForest and Boosting in a particular research topic, and adopted the best predictive model to predict counterfactuals without having to rely on the parallel trends assumption.

Research topic brief summary

During the 1980s, the state of Texas lost a civil action lawsuit where a prisoner argued that the state Department of Corrections was engaging in unconstitutional practices regarding prisoners conditions. The court ruled in favor of the prisoner, and forced the state to pursue a series of settlements. Among other orders, the court placed constraints on the number of inmates allowed per cells. Given this constraint, state legislators approved a billion dollar prison construction project that ended up doubling the state's capacity within 3 years.

Cunningham (2020) argues that the nature of this expansion allows us to use it as a natural experiment to estimate the effect of prison expansion on incarceration. He uses the synthetic control method to predict counterfactuals as in Abadie et al. (2010) by searching for the set of weights that generate the best fitting convex combination of the control units, being the best the one that minimizes root mean square error in the pre-treatment period.

His preliminary results indicate that an increase in state prison capacity caused an increase in black male incarceration. Bearing this in mind, we used a set of alternative methods learned in class to estimate counterfactuals and, therefore, measure the causal effect.

Method

In this project, we used a standard Diff-in-Diff model, and compared its results with the simple difference in outcomes predicted by the alternative method that yields the best out of sample predictive power among multiple train-test splits.

With some evidence that the Diff-in-Diff assumptions might not hold, a prominent supervised learning modelling method hopefully might predict counterfactuals with more precision and produce more robust and accurate results. Bearing the characteristics of our dataset in mind, we decided to conduct iterative model selection and utilize regularization based methods to identify the best working model.

This dataset contains observations of various prisons across all the states in the country from 1985 to 2000. So the variable year is continuous over the 15 year period. With the races primarily divided into black and white, this dataset deals with various factors that could possibly contribute to crime and the share of African American population in the state of Texas. In light of this, variables like alcohol consumption per capita (alcohol), aids mortality (aidscapita), average household income (income), unemployment rate (ur), share of the population in poverty (poverty), share of the population which is African American (black) and the share of the population which is 15 to 19 years old (perc1519) were considered. Since we were looking at increased incarceration among the black male population with prison capacity increase, number of black male prisoners (bmprison) was our dependent variable.

RESULTS

^c Differences-in-Differences

In order to have a baseline model which preserve the same parameters that were included in Cunningham's analysis, the Diff-in-Diff model is:

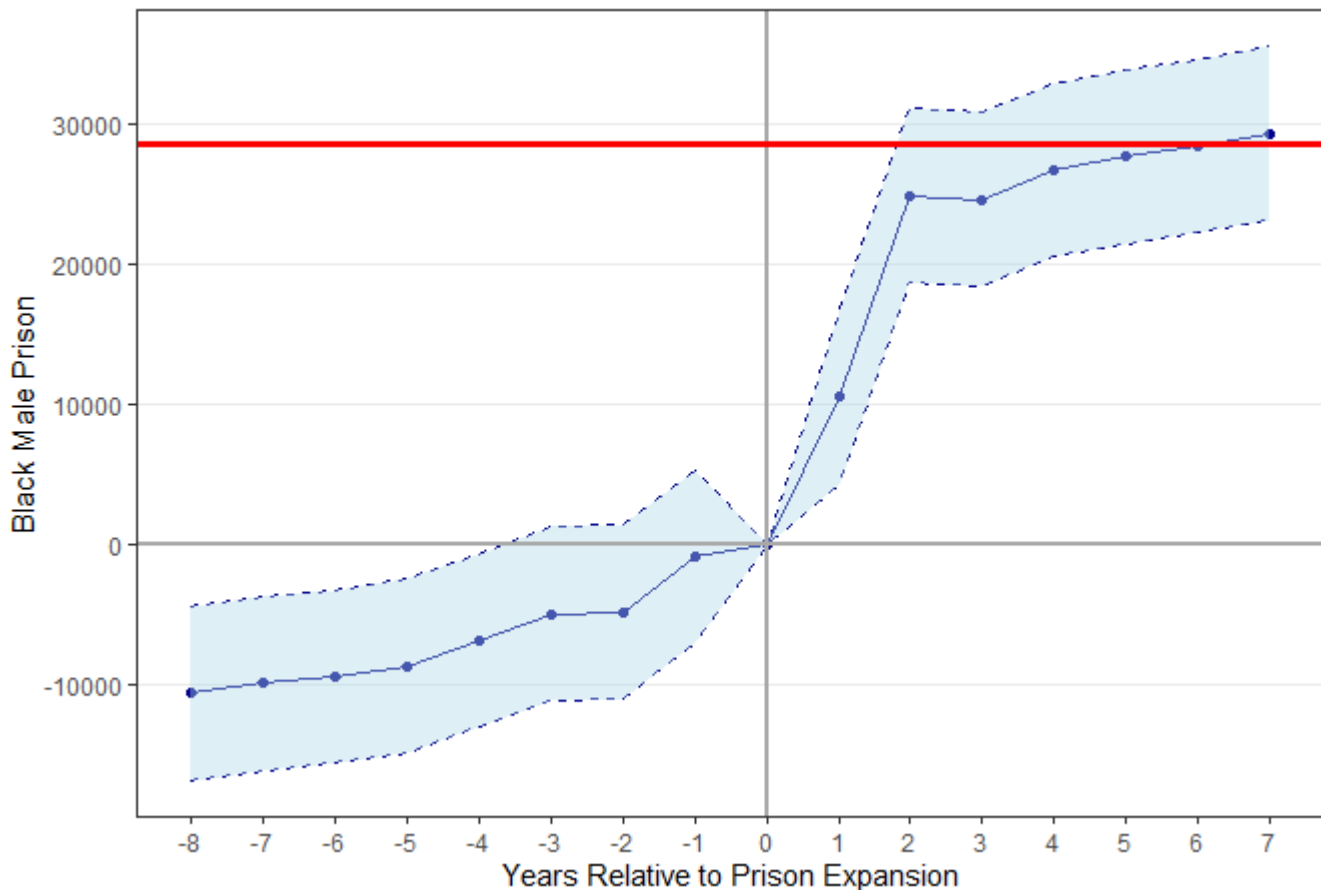
```
# bmprison ~ alcohol + aidscapita + income + ur + poverty + black + perc1519 + year +
state + year_after1993*state texas
```

The model indicates that the expansion of the state prison capacity is associated with an increase of 28,454.82 black male prisoners, holding all else fixed.

[illegible]

When decomposing the effect in each year, we get the point estimates shown in Figure 1. The coefficients capture how the treatment group differs from the control group when controlling for multiple factors and when considering state and year fixed effects. It also allows us the test the plausibility of parallel trends in the pre-treatment period. As we are including controls and fixed effects, there should be less to be explained by the coefficients to the left of the grey vertical line since the only difference should be the treatment itself, and it didn't occur in years prior to the intervention.

Figure 1: The effect decomposition



A test of joint significance of the leads coefficients, as in Kearney and Levine (2015), reject the null hypothesis that they are jointly equal to zero (p-value = 0.006). Therefore, it provides evidence that the parallel trends assumption doesn't hold even in the pre-treatment period, indicating the necessity of exploring different methods.

```
##
## Table 3: The test of joint significance
## =====
## Statistic Res.Df  Df   Chisq  Pr(> Chisq)
## -----
## Mean          731.500  7.000  19.959    0.006
## -----
```

Therefore, we tested 3 alternative models to find out the best predictive one other than assuming the parallel trends. As mentioned before, we used Lasso regression, RandomForest, and Boosting model respectively and tested their performances by K-fold validation.

↳ Lasso Regression

First, we fit a lasso regression. From the baseline model we used in diff-in-diff analysis, we added one more variable - 'crack', hoping it can enhance our model's predictive power, and considered all possible interactions.

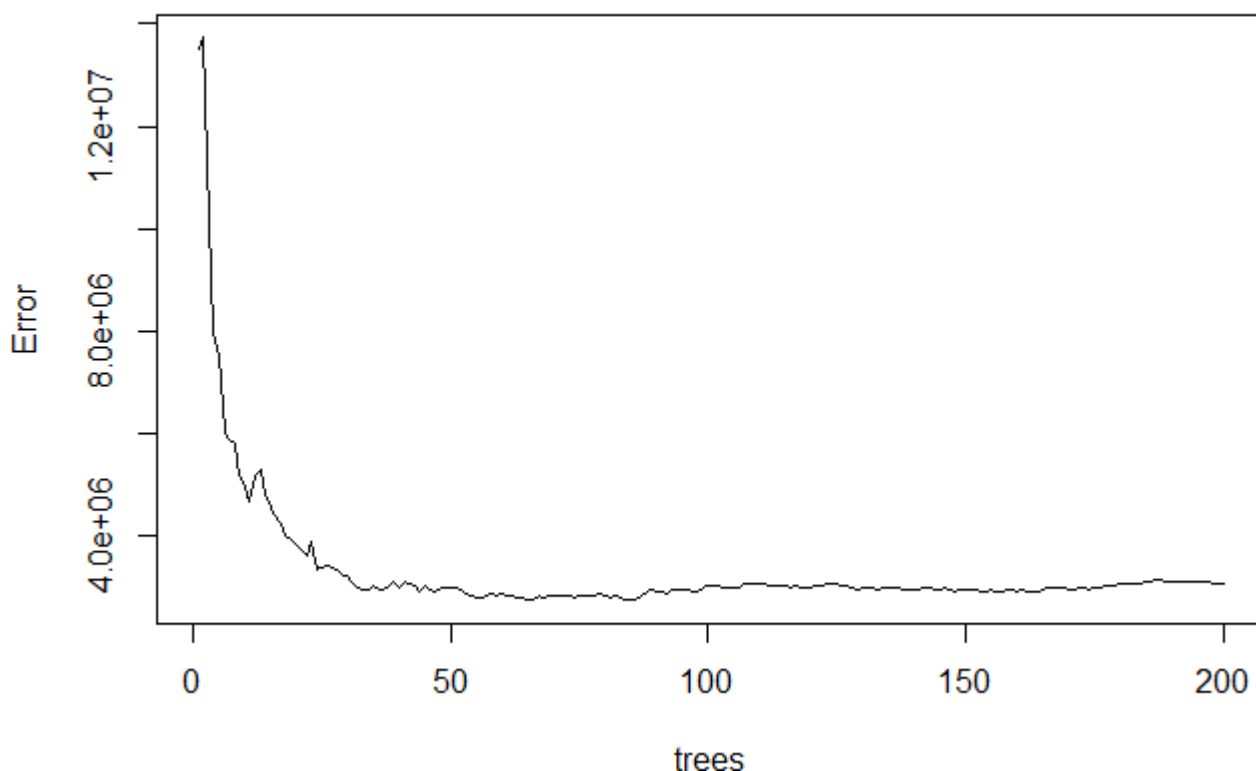
As a result, we obtained a model with 181 variables with an intercept. Then we did K-fold cross validation to check RMSE when K is 10. We used a train-test split and repeated the step from 1 to K repetitions by running a loop. Our train set and test set were both subsets of our whole dataset except the observations from the state of Texas after 1993, which is what we want to predict. By doing it, we can measure how the model estimate the change of black male prisoners which is not affected by the policy implementation.`

When we calculate RMSE for the backward selection model, it turned out to be 408.42.

↳ RandomForest

After this, we fit a RandomForest model and also did K-fold cross validation with the same baseline model we used in our lasso regression above. We started with 200 trees and as Figure 2 shows 200 is enough to reduce our errors.

Figure 2: Number of trees vs error in Randomforest



The K-fold validation result shows that the RMSE is 1641.85 which is about 4 times larger than the RMSE of lasso regression.

⁝ Boosting

Lastly, we used a boosting model with the same baseline model and did K-fold validation as we did above.

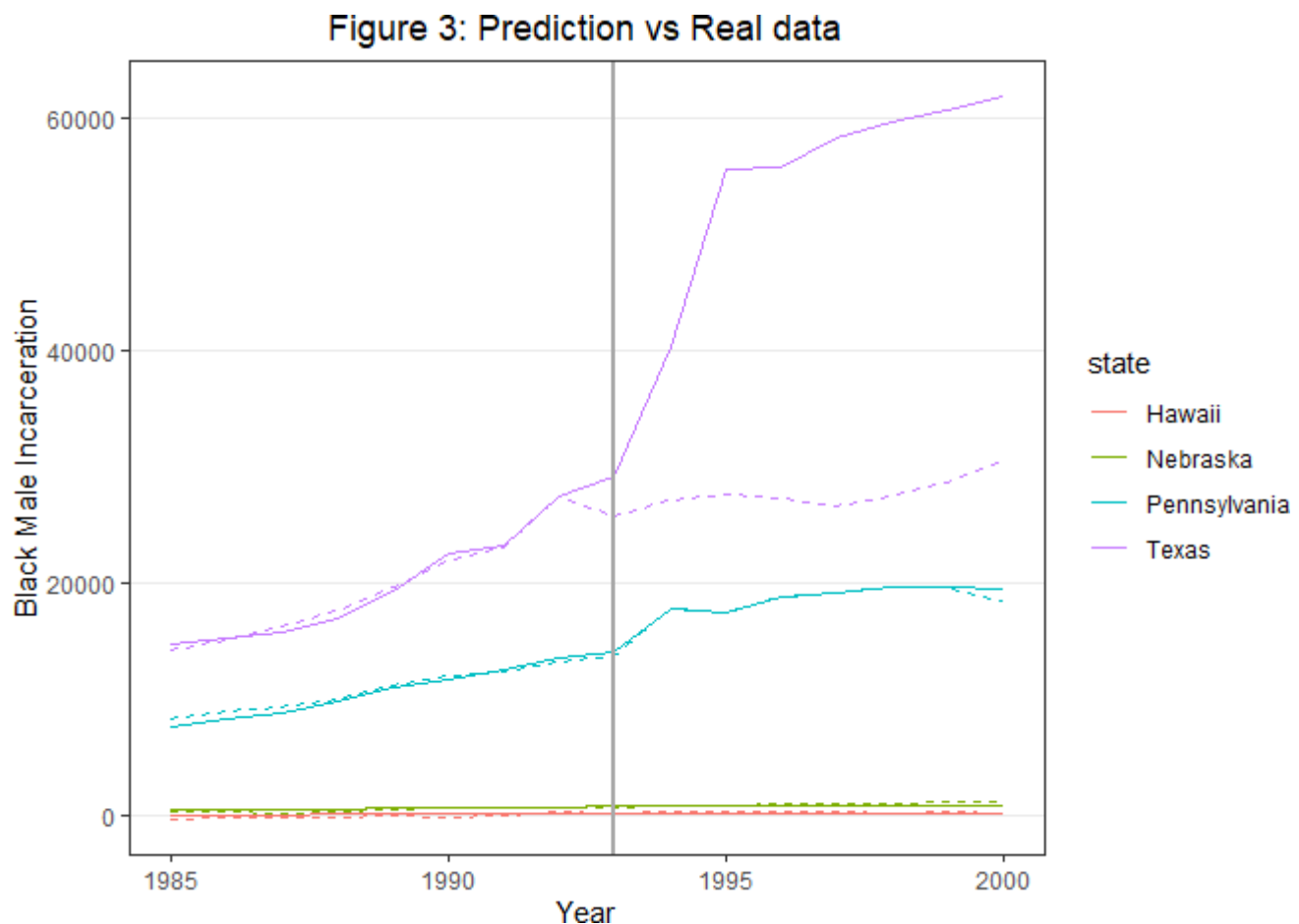
The result of our K-fold cross validation shows that the RMSE is 866.89 which is lower than the RandomForest model but still higher than the lasso regression. [Table 3] shows that the lasso regression has the best predictive power among all the models that we tested.

Table 4: RMSE results for each model

```
##
##
## +-----+-----+-----+-----+
## | **Model** | Lasso | Randomforest | Boosting |
## +-----+-----+-----+-----+
## | **RMSE** | 408.42 | 1641.85 | 866.89 |
## +-----+-----+-----+-----+
```

⁝ Comparing the best model's predictions with the observed data

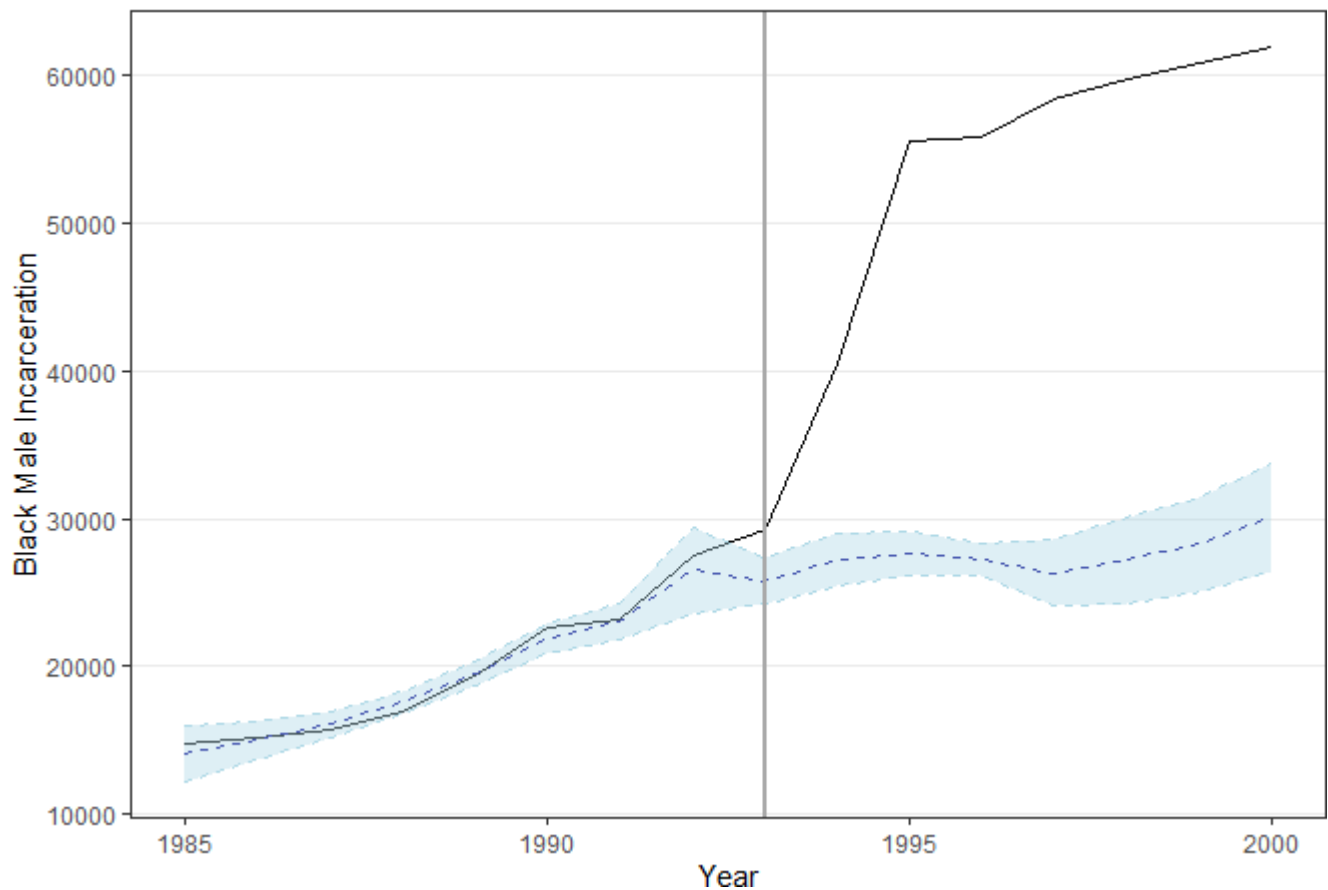
Since we have assessed our best predictive model, now we can compare its predictions with the real data in our whole data set. We can see how our prediction goes along with the real data in Figure 3. It shows the change of black male incarceration in the treated state, Texas, with 5 randomly chosen states.



In Figure 3, we can see two interesting findings. One is that Texas is showing clearly different movement from our predicted trend after the treatment in 1993. The other is that Our prediction from the lasso model fits very well on real data of controlled states.

For inference purposes, it is recommended to estimate a confidence interval rather than showing the point estimate only. Therefore, we used a bootstrap to calculate the standard deviation of the parameter's resampling distribution, and illustrated the results on figure 4.

Figure 4: Estimating a confidence interval



The average treatment effect calculated by taking the average of the simple difference of outcomes between the observed data and the predictions yielded by our most accurate model was:

```
## [1] 25254.73
```

When comparing it to the Diff-in-Diff estimator, we can observe a decrease of approximately 12% on the magnitude of the effect.

⁹ Conclusion

The analysis showed that alternative supervised learning methods can play a big role in predicting counterfactuals either when there are reasons to believe that the traditional assumptions don't hold, or to corroborate the assumptions as well. It is important to notice that it is upon to the researcher's discretion how to do it in practice, and it might open up space for "p-hacking" when moving away from the best practices. In that sense, peer review/validation is crucial to ensure that the predictions are being yielded by models that minimize out of sample root mean square error, and randomness is fundamental to guarantee that the results aren't being conveniently tampered.

Given that our dataset was consisted of only 816 observations and a limited amount of covariates, our best judgement is that the alternative model yielded satisfactory results. We could have increased our model's predictive power if we had had more observations, however, our model fits very well on real data of controlled states suggesting the possibility of predicting counterfactuals.

⁹ References

Wooldridge, J.M. (2013). Introductory econometrics: A modern approach

Angrist, J and Pischke, J.S. (2018). Mostly Harmless Econometrics

Cunningham, Scott (2020). CAUSAL INFERENCE: THE MIXTAPE

Abadie, A, Diamond, A and Hainmueller, J (2010). Synthetic Control Methods for Comparative Case Studies: Estimating the Effect of California's Tobacco Control Program

Kearney, M.S. and Levine, P.B. (2015) Media Influences on Social Outcomes: The Impact of MTV's 16 and Pregnant on Teen Childbearing