Nihal Temüge
12318000

Empirical Project

Machine Learning in Econometrics

Introduction

Double Machine Learning (Double ML) represents a cutting-edge approach in the field of causal inference, seamlessly integrating machine learning algorithms with the augmented inverse propensity score weighting (AIPW) method for estimating causal treatment effects in observational studies. This powerful framework is employed in our empirical project to assess the effectiveness of new German policies aimed at reducing gender inequality in wages and employment.

In observational studies, researchers frequently grapple with the challenge of estimating causal effects when treatments are not randomly assigned, resulting in potential selection biases. To address this issue, propensity score weighting is commonly employed. The propensity score represents the likelihood of receiving treatment given a set of observed covariates, and inverse propensity score weighting involves reweighting the data to create a pseudo-population that closely resembles a randomized experiment.

Double ML builds upon AIPW by introducing two machine learning models: one for estimating the propensity scores (referred to as the "propensity score model") and another for estimating the treatment effect (referred to as the "response model"). By leveraging machine learning models for both steps, Double ML can effectively handle intricate relationships between covariates and treatment assignment or outcome. This is especially advantageous when working with high-dimensional data.

In the context of policy evaluation, it is paramount to assess the impacts of public interventions on various social, economic, and environmental challenges. Traditionally, researchers have relied on estimating the overall average treatment effect to evaluate policy success. However, this approach fails to capture the heterogeneity in treatment effects among different subgroups within the population. As a consequence, policymakers might miss crucial insights and disparities, leading to less-than-optimal policy decisions.

To overcome these limitations, our project advocates for the adoption of CATE analysis (Conditional Average Treatment Effect) as the primary focus in policy evaluation. CATE is crucial because it provides more nuanced insights into how a particular policy or intervention affects diverse subgroups within the population. By relying solely on methods that consider the overall average treatment effect, policymakers may overlook significant variations in treatment effects among different groups. This, in turn, can lead to suboptimal policy decisions and missed opportunities for implementing more targeted and effective interventions. Therefore, embracing CATE analysis offers a more comprehensive and informed approach to policy evaluation. Here are some reasons why CATE is important for policy evaluation:

- **Identifying Differential Effects:** CATE allows policymakers to identify which subgroups of the population benefit the most from a policy and which groups might not experience significant improvements. This information helps target resources to those who need them the most, reducing wastage and maximizing the policy's overall impact.
- **Equity and Fairness:** By analyzing CATE, policymakers can assess whether a policy has disproportionate effects on certain demographic groups, income levels, or regions. If the policy benefits one group significantly more than others, it may raise concerns about equity and fairness, prompting policymakers to make adjustments to address these disparities.

Nihal Temüge
12318000

- **Cost-Effectiveness:** Understanding CATE can help policymakers determine the cost-effectiveness of a policy. By focusing resources on groups where the policy has the most substantial impact, the overall cost-effectiveness of the intervention can be improved.
- **Policy Tailoring:** CATE analysis can guide policymakers in tailoring policies to specific subgroups. By understanding which factors influence the effectiveness of the policy, policymakers can design more targeted and tailored interventions for different groups.
- **Unintended Consequences:** Policies can sometimes have unintended consequences, and these effects may not be immediately apparent when considering the overall average treatment effect. CATE analysis can help identify such unintended consequences in specific subgroups, allowing policymakers to mitigate any adverse effects.
- **Generalization of Findings:** CATE analysis provides more robust findings by accounting for heterogeneity in treatment effects. This allows policymakers to generalize the policy's impact to specific subgroups more accurately and make informed decisions about scaling up or adapting the intervention to different contexts.
- **Policy Iteration and Improvement:** CATE analysis can provide valuable feedback on the policy's performance, allowing for policy iteration and continuous improvement. Policymakers can use CATE estimates to refine and enhance the intervention over time.

Hence, with the conditional independence assumption i.e. $D_i \perp (Y_{0i}, Y_{1i}) \mid W_i$ for some (potentially high-dimensional) vector of observables W_i, Conditional Average Treatment Effects (CATEs) are defined for some(low-dimensional) vector of observables X_i, as follows:

$$CATE(x) := E[Y_{1i} - Y_{0i} \mid X_i = x]$$

With pre-specified covariates X_i, first pseudo outcomes Ψ_i is constructed using ML method random forest , then low-dimensional regression of Ψ_i on $X_i \rightarrow$ predicted value at x is the estimator of CATE(x). Precise estimates of the Conditional Average Treatment Effect (CATE(x)) can be achieved with confidence when certain conditions are met. Specifically, when the dimensionality of X_i (i.e., W_i) is low, and a sufficient number of observations exist in each leaf, we can use shallow trees to ensure accurate estimation. In this scenario, the square root of the sample size multiplied by the difference between CATE(x) and its estimate approaches a normal distribution with mean 0 and variance V(x) (where V(x) is greater than 0). Employing these strategies allows us to obtain reliable and high-quality estimates of CATE(x).[1]

[1] ML in Econometrics Lecture notes: Machine Learning in Econometrics - Heterogeneous Treatment Effects (lmu.de)

Nihal Temüge
12318000

Methodology

1) Variable Selection via LASSO Model:

The first step of the analysis is conducting the LASSO model for variable selection. Firstly, a set of assumptions should be satisfied for the LASSO model: For example, it requires s log(max{p, n})/n to be small, and also it requires sparsity which means only a relatively small number of regressors are important. Additionally, LASSO typically performs well when the sample size is larger than the number of predictors.[2] Hence, I first run a LASSO model.After fitting the LASSO model via "glmnet" package in R, the optimal value of the lamba (penalty parameter) that minimizes the mean squared error (MSE) is obtained. LASSO model with the optimal lamba is fitted and then selected variables and their coefficients are constructed.

2) Double ML for Augmented Inverse Propensity Score Weighting (AIPW)

In this step, selected variables are used for the analysis, and methods in the" Double machine learning based program evaluation under unconfoundedness" paper by Knaus, M. C. (2022) are followed. Firstly, the nuisance parameters are estimated with Random Forest[3]. Honest splitting in the "grf" R-package[4] and 5-fold cross-fitting are implemented. To obtain estimates of μˆ(w, x), I conduct separate outcome regressions for each treatment group. Additionally, I estimate propensity scores individually for each treatment by employing a random forest with the treatment indicator as the outcome variable. Subsequently, the calculated propensity scores are normalized to ensure they sum up to one within each individual.

First, Group Treatment Effects (GATES) are estimated for subgroups by gender. The Double Machine Learning (DML) approach for average treatment effects can be implemented efficiently using a standard OLS regression with the pseudo-outcome[5] and dummy variables for treatment groups. This yields reliable estimates of average treatment effects in my analysis. Pseudo-outcome is constructed as follows:

$$Y\sim ATE=m\hat{}(1,X)-m\hat{}(0,X)+ W(Y-m\hat{}(1,X))/e\hat{}(X)-(1-W)(Y-m\hat{}(0,X))/1-e\hat{}(X),$$

where $m\hat{}(1,X)-m\hat{}(0,X)$ is outcome predictions and

$W(Y-m\hat{}(1,X))/e\hat{}(X)-(1-W)(Y-m\hat{}(0,X))/1-e\hat{}(X)$ is weighted residuals.[6]

Additionally, the analysis extends to estimate nonparametric Conditional Average Treatment Effects (CATEs) for continuous variables such as age, and years of education providing valuable insights into how the treatment effects vary across different subgroups based on these covariates. Overall, this approach enables a comprehensive evaluation of policy effectiveness and treatment impacts.

---

[2] ML in Econometrics Lecture notes: Machine Learning in Econometrics - LASSO Theory (lmu.de)
[3] Breiman, L. (2001). Random forests. Machine Learning, 45 (1), 5–32.
[4] Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. Annals of Statistics, 47 (2), 1148 - 1178.
[5] Knaus, M. C. (2022). Double machine learning based program evaluation under unconfoundedness, *The Econometrics Journal*, forthcoming, arXiv
[6] Knaus, M. C. (2022). Double machine learning based program evaluation under unconfoundedness, *The Econometrics Journal*, forthcoming, arXiv

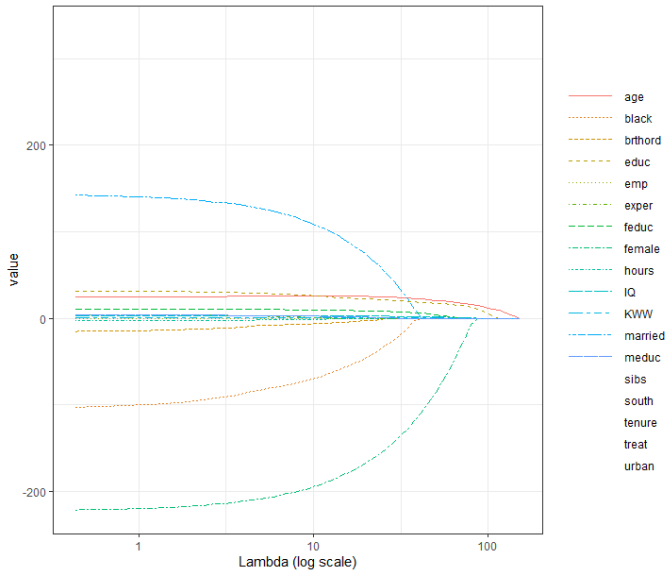
Nihal Temüge
12318000

Results of The Analysis

a) Results For Wages

1. LASSO Model

Before creating Augmented Inverse Propensity Score Weighting (AIPW) model for policy evaluation, I developed LASSO analysis for variable selection. Firstly, "wage" is assigned as the outcome variable and the predicted matrix is created by using all variables except wage, id, and year. After fitting the LASSO model, I identified the optimal lambda value (penalty parameter) that minimizes the mean squared error (MSE). Secondly, the LASSO model is fitted by using the optimal lambda value which is found as 2.114331 in the previous step and selected variables which means those have non-zero coefficients are obtained. The selected variables (i.e. having lines in the legend) can be seen in the following figure below:



According to the results, it might be argued that age, education, married, and female are the most important explanatory variables for wages as well as a couple of other variables.

2. Double ML for Augmented Inverse Propensity Score Weighting (AIPW)

In this project, Augmented IPW is used to extract the treatment effect heterogeneity since Inverse Propensity Score Weighting has a regulation bias. This analysis is useful to target treatment to those who benefit the most and introduce good policies.

Firstly, I created the pseudo-outcome by using the "DML_aipw" function defined in the "causalDML" package in R. With this package, I developed 5-fold cross-fitting. The results are as follows:

	ATE	SE	t	p
1 – 0	0.287855	0.039551	7.2781	5.81e-13 ***

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Nihal Temüge
12318000

The results above show that the Average Treatment Effect (ATE) of the policy on wages is 0.287855. The next step of the analysis is creating the forest and tuning it with the DML_aipw function. The pseudo-outcome can now be used to estimate different heterogeneous effects. I used standard regression models but by using the pseudo-outcome instead of a real outcome I model effect size and not outcome level. To check gender differences, I split the sample by gender and rerun the whole analysis in the subsamples separately. With the pseudo-outcome stored in aipwATEdelta this boils down to running an OLS regression with the female indicator as a single regressor.

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	0.0943	0.0685	1.377	1.689e-0
female	0.4094	0.07655	5.348	1.049e-07

Multiple R-squared: 0.02087, Adjusted R-squared: 0.02012, F-statistic: 28.6 on 1 and 1311 DF, p-value: 1.049e-07

The table above shows the effect size differences of the policy between groups i.e., females and males. This means the intercept gives us the average of the reference group (men) and the coefficient tells us how much higher the effect is for women. In this case, I find significant gender differences in log of wages.

I run an OLS regression without constant to obtain the gender-specific effect instead of differences between groups and I find the followings:

	Estimate	Std. Error	t value	Pr(>|t|)	CI Lower	CI Upper	DF
male	0.0943	0.06851	1.377	1.689e-0	-0.04009	0.2287	1311
female	0.5037	0.03415	14.748	1.169e-45	0.43667	0.5707	1311

Multiple R-squared: 0.06158, Adjusted R-squared: 0.06015, F-statistic: 109.7 on 2 and 1311 DF, p-value: < 2.2e-16

3. Best Linear Prediction

The next step is developing BLP analysis by using the pseudo-outcome. Hence, it can be completely agnostic about the outcome model and receive a nice summary of the underlying effect heterogeneity. OLS results are as follows:

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	2.063127	0.632655	3.26106	1.139e-03
Xhours	0.011176	0.010145	1.10162	2.708e-01
Xfemale	0.420547	0.074274	5.66208	1.840e-08
XIQ	-0.009752	0.003776	-2.58223	9.925e-03
XKWW	0.002829	0.006378	0.44356	6.574e-01
Xeduc	-0.004363	0.032721	-0.13335	8.939e-01
Xexper	0.008081	0.011452	0.70561	4.806e-01
Xtenure	-0.024079	0.010805	-2.22853	2.602e-02
Xage	-0.039311	0.011702	-3.35942	8.038e-04
Xmarried	0.155601	0.167910	0.92669	3.543e-01
Xblack	-0.003822	0.129349	-0.02955	9.764e-01
Xsouth	-0.129949	0.096654	-1.34447	1.790e-01
Xurban	-0.043782	0.089088	-0.49144	6.232e-01
Xsibs	0.011410	0.016912	0.67466	5.000e-01
Xbrthord	-0.017702	0.022648	-0.78160	4.346e-01
Xmeduc	-0.019772	0.022014	-0.89815	3.693e-01
Xfeduc	0.013458	0.017486	0.76964	4.417e-01

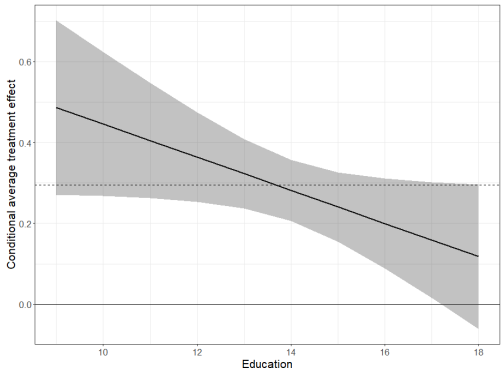
Nihal Temüge
12318000

According to the results, being female increases the effect of the policy on log of wages by 0.42. One unit increase in result on an IQ test slightly decreases the effect of the policy on log of wages by 0.01. Additionally, being one year older decreases the effect by 0.04.

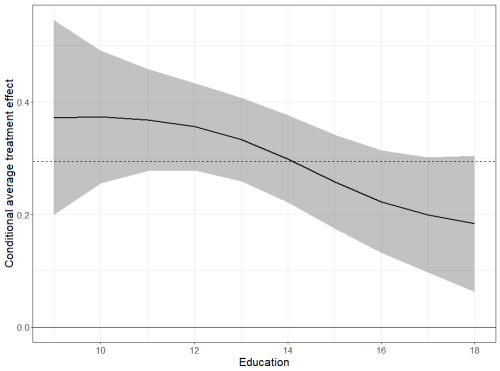
4. Non-parametric heterogeneity

By using pseudo-outcome again, I also run non-parametric regressions to estimate heterogeneous effects which means that the analysis is also agnostic regarding the functional of effect heterogeneity with the outcome and propensity score models.[7] It`s important, especially for continuous variables. Hence, I run spline regressions and Kernel regressions for ages and education.

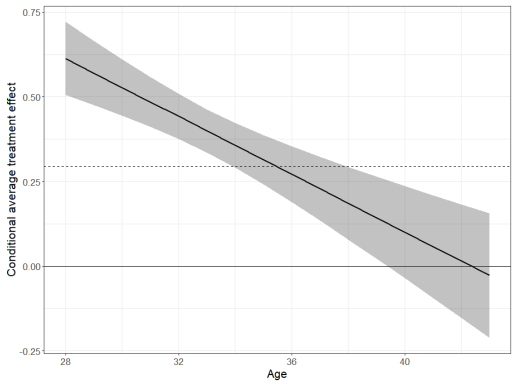
Spline Regression



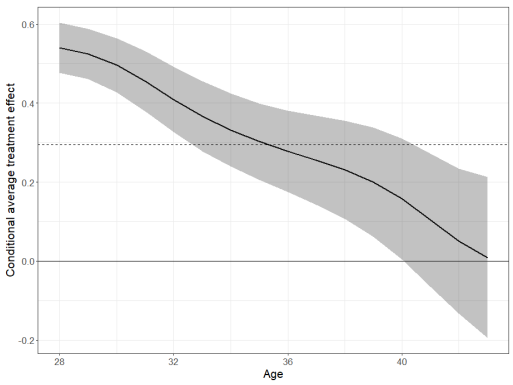
Kernel Regression



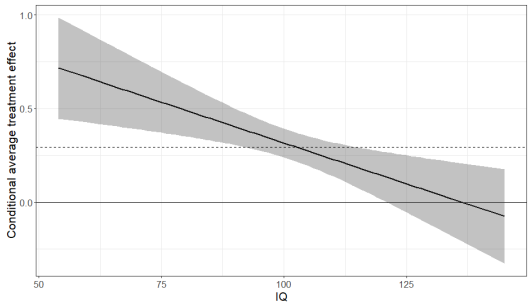
Spline Regression



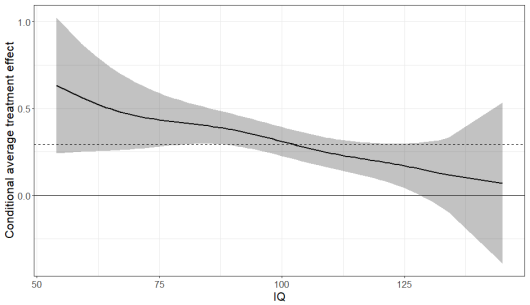
Kernel Regression



Spline Regression



Kernel Regression



[7]Knaus, M. C. (2022). Double machine learning based program evaluation under unconfoundedness, *The Econometrics Journal*, forthcoming, arXiv

Nihal Temüge
12318000

Graphs of Spline Regressions and Kernel regressions for education and age show that the effect of the policy on log of wages decreases slightly with an increase in years of education, especially after 12 years of education it decreases more until 17 years of education, then it slows down. That means less educated people benefit more from the policies. Similarly, the effect of the policy decreases with increases in age and results in an IQ test. Hence, younger people benefit more than older ones. Similarly, people with lower IQ test results benefit more than those who have higher IQ test results.

b) Results For Employment

I implement the same analysis to estimate the effect of the policies on employment. Firstly, I run the LASSO model to select important variables. Then, I obtain them by running LASSO model again with the optimal lambda value as before. I created the pseudo-outcome by using the "DML_aipw" function and developed 5-fold cross-fitting. The results are as follows:

	ATE	SE	t	p
1 – 0	- 0.0049596	0.0029431	- 1.6852	0.09216 .

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

The results above show that the Average Treatment Effect (ATE) of the policy on employment is -0.0049596. The next step of the analysis is creating the forest and tuning it with the DML_aipw function. I run standard regression models with the pseudo-outcome to estimate different heterogeneous effect sizes. To check gender differences, I split the sample by gender and rerun the whole analysis in the subsamples separately. With the pseudo-outcome stored in aipwATEdelta this boils down to running an OLS regression with the female indicator as a single regressor.

	Estimate	Std.Error	t value	Pr(>|t|)
(Intercept)	-0.0004781	0.004253	-0.1124	0.9105
female	-0.0022885	0.004611	-0.4963	0.6198

I find no significant gender differences in the effect of policies on employment. Additionally, I run an OLS regression without constant to obtain the gender-specific effect instead of differences between groups and I find the followings:

	Estimate	Std.Error	t value	Pr(>|t|)	CI Lower	CI Upper	DF
male	-0.0004781	0.004253	-0.1124	0.9105	-0.008821	0.0078648	1470
female	-0.0027666	0.001782	-1.5528	0.1207	-0.006261	0.0007283	1470

Multiple R-squared: 0.000482,Adjusted R-squared: -0.0008779 , F-statistic: 1.212 on 2 and 1470 DF, p-value: 0.2979

The next step is developing BLP analysis by using the pseudo-outcome. According to the OLS results, none of the variables are effective for change in the size of the effect of the policy on employment.

Additionally, I run Spline and Kernel regressions for Age, Education, and IQ again but I couldn`t find significant results for the different effects of the policies on employment.[8]

[8] See Appendix

Nihal Temüge
12318000

Potential Difficulties

The effectiveness of a policy is influenced by a range of factors that differ between countries like Germany and the United States. Cultural and social differences play a crucial role, as perceptions and responses to gender-related policies may vary significantly based on the distinct cultural norms and social contexts of each nation. Moreover, the legal framework in Germany, encompassing its unique labor laws and regulations, might diverge from that of the U.S., potentially impacting the implementation and effectiveness of similar gender equality policies. Furthermore, variations in the composition of industries and firms in Germany compared to the U.S. could lead to different patterns of gender inequality within the workforce. Another vital aspect to consider is the availability and quality of data in the German context, which may differ, potentially affecting the accuracy and generalizability of research findings related to gender disparities. Lastly, the success of gender equality policies relies not only on their design but also on how effectively they are implemented and enforced, making policy implementation a crucial factor in addressing gender inequality in both countries. Understanding these various elements is imperative for crafting effective and targeted policies aimed at promoting gender equality in diverse national contexts:

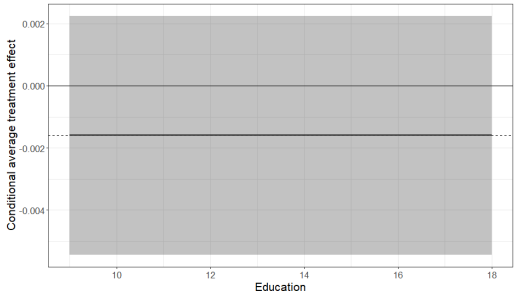
Conclusion

In conclusion, the analysis of the data reveals insightful patterns regarding the impact of the policy on wages across genders. The findings indicate that being female has a positive effect, increasing the policy's influence on log of wages by 0.42. Conversely, a one-unit increase in IQ test results appears to have a slight dampening effect, reducing the policy's impact on log of wages by 0.01. Moreover, the age factor plays a role, with every additional year of age leading to a decrease in the policy's effect by 0.04. The graphs of Spline Regressions and Kernel regressions further elucidate the relationship between education, age, and the policy's impact. It is evident that the effect on log of wages declines slightly with increased years of education, especially after 12 years, and continues to decrease until 17 years, where it slows down. This suggests that individuals with lower levels of education benefit more from the policies. Similarly, the policy's effect diminishes as age increases, indicating that younger individuals benefit more than their older counterparts. Additionally, those with lower IQ test results experience a greater positive impact from the policy compared to individuals with higher IQ test results. However, I did not observe any statistically significant gender differences in the effect of policies on employment. These nuanced insights provide valuable guidance for policymakers to tailor interventions more effectively, ensuring that the policies are targeted towards specific demographic groups for greater overall effectiveness in reducing gender inequality in wages.

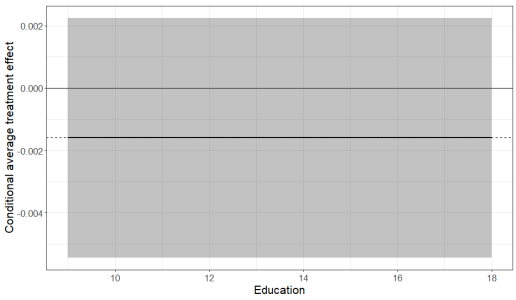
Nihal Temüge
12318000

Appendix

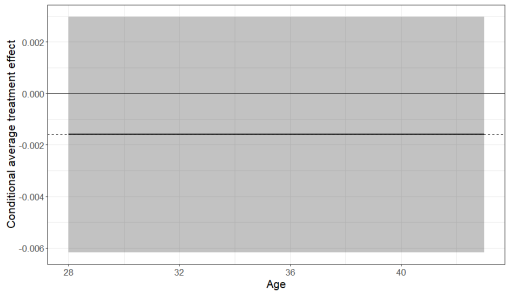
Spline Regression



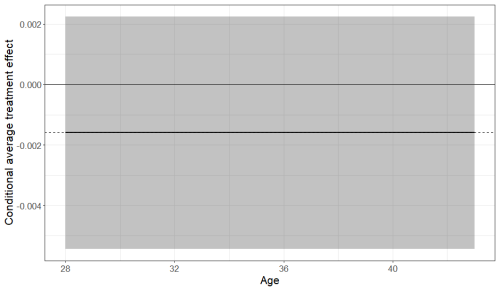
Kernel Regression



Spline Regression



Kernel Regression



References

1) Athey, S., Tibshirani, J., & Wager, S. (2019). Generalized random forests. Annals of Statistics, 47 (2), 1148 - 1178.
2) Breiman, L. (2001). Random forests. Machine Learning, 45 (1), 5–32.
3) Heiler, P., Knaus, M. C. (2021). Effect or Treatment Heterogeneity? Policy Evaluation with Aggregated and Disaggregated Treatments, arXiv
4) Knaus, M. C. (2022). Double machine learning based program evaluation under unconfoundedness, *The Econometrics Journal*, forthcoming, arXiv
5) ML in Econometrics Lecture notes: Machine Learning in Econometrics - Heterogeneous Treatment Effects (lmu.de)
6) ML in Econometrics Lecture notes: Machine Learning in Econometrics - LASSO Theory (lmu.de)

Confirmation

I confirm that this report is based on my own work. In preparing this report, I have not received any help from another human nor have I discussed any aspects of the empirical project with others.