Modified Causal Forest: Estimation

Section 1: General information

Welcome to the mcf estimation and optimal policy package.
This report provides you with a summary of specifications and results. More detailed information can be found in the respective output files. Figures and data (in csv-format, partly to recreate the figures on your own) are provided in the output path as well.

Output information for MCF ESTIMATION
All outputs: output_treatment_effect
Subdirectories with figures and data are named ate_iate, gate, and common support and contain the content related to their name.
Detailed text output: output_treatment_effect/txtFileWithOutput.txt
Summary text output: output_treatment_effect/txtFileWithOutput_Summary.txt

BACKGROUND

ESTIMATION OF EFFECTS
The MCF is a comprehensive causal machine learning estimator for the estimation of treatment effects at various levels of granularity, from the average effect at the population level to very fine grained effects at the (almost) individual level. Since effects at the higher levels are obtained from lower level effects, all effects are internally consistent. Recently, the basic package has been appended for new average effects as well as for an optimal policy module. Effect estimation is implemented for identification by unconfoundedness as well as by instrumental variables. While unconfoundedness estimation can deal with multiple treatments, instrumental variable estimation is restricted to binary instruments and binary treatments.The basis of the MCF estimator is the the causal forest suggested by Wager and Athey (2018). Their estimator has been changed in several dimensions which are described in Lechner (2018). The main changes relate to the objective function as well as to the aggreation of effects. Lechner and Mareckova (2024) provide the asymptotic guarantees for the MCF and compare the MCF, using a large simulation study, to competing approaches like the Generalized Random Forest (GRF, Athey, Tibshirani, Wager, 2019) and Double Machine Learning (DML, Chernozhukov, Chetverikov, Demirer, Duflo, Hansen, Newey, Robins, 2018, Knaus, 2022). In this comparison the MCF faired very well, in particular, but not only, for heterogeneity estimation. Some operational issues of the MCF are discussed in Bodory, Busshof, Lechner (2022). There are several empirical studies using the MCF, like Cockx, Lechner, Boolens (2023), for example.

References
- Athey, S., J. Tibshirani, S. Wager (2019): Generalized Random Forests, The Annals of Statistics, 47, 1148-1178.
- Athey, S., S. Wager (2019): Estimating Treatment Effects with Causal Forests: An Application, Observational Studies, 5, 21-35.
- Bodory, H., H. Busshoff, M. Lechner (2023): High Resolution Treatment Effects Estimation: Uncovering Effect Heterogeneities with the Modified Causal Forest, Entropy, 24, 1039.
- Bodory, H., F. Mascolo, M. Lechner (2024): Enabling Decision Making with the Modified Causal Forest: Policy Trees for Treatment Assignment, Algorithm, 17, 318.
- Chernozhukov, V., D. Chetverikov, M. Demirer, E. Duflo, C. Hansen, W. Newey, J. Robins (2018): Double/debiased machine learning for treatment and structural parameters, Econometrics Journal, 21, C1-C68.

Modified Causal Forest: Estimation

- Cockx, B., M. Lechner, J. Bollens (2023): Priority to unemployed immigrants? A causal machine learning evaluation of training in Belgium, Labour Economics, 80, Article 102306.
- Knaus, M. (2022): Double Machine Learning based Program Evaluation under Unconfoundedness, Econometrics Journal.
- Lechner, M. (2018): Modified Causal Forests for Estimating Heterogeneous Causal Effects, arXiv.
- Lechner, M. (2023): Causal Machine Learning and its Use for Public Policy, Swiss Journal of Economics & Statistics, 159:8.
- Lechner, M., J. Mareckova (2024): Comprehensive Causal Machine Learning, arXiv.
- Lechner, M., J. Mareckova (2025): Comprehensive Causal Machine Learning with Instrumental Variables, mimeo.
- Wager, S., S. Athey (2018): Estimation and Inference of Heterogeneous Treatment Effects using Random Forests, Journal of the American Statistical Association, 113:523, 1228-1242.

Modified Causal Forest: Estimation

Section 2: MCF estimation

METHOD
Standard MCF method used. Nearest neighbour matching performed using the Prognostic Score.
Feature selection not is used.
Local centering is used.
Common support is enforced.

VARIABLES
Outcomes: sal_avg, sal_3, sal_4, sal_5, sal_6, sal_7, sal_8, sal_9, empl_ttl, empl_chge
Treatment: ptype (with values 0 1 2)
Ordered confounders: age, sex, school, voc_deg, reg_al, reg_ser, reg_pro, reg_agri, sect_al, prof_al, unem_x0, olf_x0, empl_x0, earn_x0, emplx1_1, emplx1_2, emplx1_3, emplx1_4, emplx2_1, emplx2_2, emplx2_3, emplx2_4, earnx1, earnx2, lmp_cw, sex, age, school, voc_deg, lmp_cw
Unordered (categorical) confounders: nation, region, nation
Ordered heterogeneity variables (few values, continuous variables are discretized): sex, age, school, voc_deg, lmp_cw
Unordered heterogeneity variables: nation

EFFECTS ESTIMATED
Average Treatment Effect (ATE), Group Average Treatment Effect (GATE), Individualized Average Treatment Effect (IATE)

NOTE on unordered variables:
One-hot-encoding (dummy variables) is not used as it is expected to perform poorly with trees: It may lead to splits of one category versus all other categories. Instead the approach used is analogous to the one discussed in Chapter 9.2.4 of Hastie, Tibshirani, Friedmann (2013), The Elements of Statistical Learning, 2nd edition.

Section 2.1: MCF Training

Training uses 10 CPU cores.

Section 2.1.1: Preparation of training data (mcf training)

METHOD
Variables without variation are removed.
Variables that are perfectly correlated with other variables are removed.
Dummy variables with less than 10 observations in the smaller group are removed.
Rows with any missing values for variables needed for training are removed.

Modified Causal Forest: Estimation

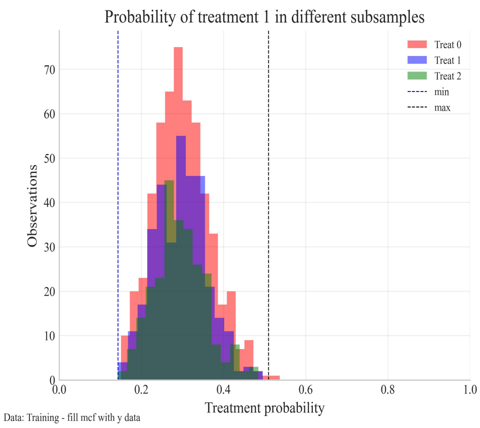
RESULTS
No relevant variables were removed.
Sample size of training data: 2288   (no observations removed).

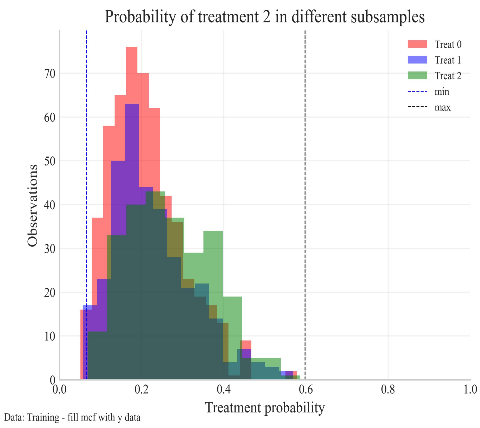
Section 2.1.2: Common support (mcf training)

METHOD
The common support analysis is based on checking the overlap in the out-of-sample predictions of the propensity scores (PS) for the different treatment arms. PSs are estimated by random forest classifiers. Overlap is operationalized by computing cut-offs probabilities of the PSs (ignoring the first treatment arm, because probabilities add to 1 over all treatment arms). These cut-offs are subsequently also applied to the data used for predicting the effects.
Overlap is determined by the min / max rule.
Cut-offs for PS are widened by 0.05.
Out-of-sample predictions are generated by 5-fold cross-validation.

RESULTS
Share of observations deleted: 1.14%
Number of observations remaining: 2262

Common support plots





Section 2.1.3: Local centering (mcf training)

METHOD
Local centering is based on training a regression to predict the outcome variable conditional on the features (without the treatment). The regression method is selected among various versions of Random Forests, Support Vector Machines, Boosting methods, and Neural Networks of scikit-learn.

Modified Causal Forest: Estimation

The best method is selected by minimizing their out-of-sample Mean Squared Error using 5-fold cross-validation. The full set of results of the method selection step are contained in output_treatment_effect/txtFileWithOutput.txt.
The respective out-of-sample predictions are subtracted from the observed outcome in the training data used to build the forest. These out-of-sample predictions are generated by 5-fold cross-validation.

RESULTS
Out-of-sample fit for Random Forest of Ey|x (R2) for sal_avg: 31.10%

Section 2.1.4: Forest

METHOD and tuning parameters
Method used for forest building is MSE & MCE Penalty "MSE of treatment variable".
The causal forest consists of 1000 trees.
The minimum leaf size is 5.
The number of variables considered for each split is 10.
The share of data used in the subsamples for forest building is 67%.
The share of the data used in the subsamples for forest evaluation (outcomes) is 100%.
Alpha regularity is set to 10%.

sal_avg_lc is the outcome variable used for splitting (locally centered).
The features used for splitting are age sex school voc_deg reg_al reg_ser reg_pro reg_agri sect_al prof_al unem_x0 olf_x0 empl_x0 earn_x0 emplx1_1 emplx1_2 emplx1_3 emplx1_4 emplx2_1 emplx2_2 emplx2_3 emplx2_4 earnx1 earnx2 lmp_cw nation region.

RESULTS
Each tree has on average 104.64 leaves.
Each leaf contains on average 7.2 observations. The median # of observations per leaf is 6.
The smallest leaves have 5 observations.
The largest leaf has 64 observations.
24.03% of the leaves were merged when populating the forest with outcomes from the honesty sample.

Modified Causal Forest: Estimation

Section 2: MCF estimation

Section 2.2: MCF Prediction of Effects

Training uses 10 CPU cores.

Section 2.2.1: Common support (mcf prediction)

Share of observations deleted: 0.87%
Number of observations remaining: 2269

Section 2.2.2: Results

GENERAL REMARKS
The following results for the different parameters are all based on the same causal forests (CF). The combination of the CF with the potentially new data provided leads to weight matrices. These matrices may be large requiring some computational optimisations, such as processing them in batches and saving them in a sparse matrix format. One advantage of this approach is that aggregated effects (ATE, GATE, BGATE) can be computed by aggregation of the weights used for the IATE. Thus a high internal consistency is preserved in the sense that IATEs will aggregate to GATEs, which in turn will aggregate to ATEs.

ESTIMATION
Weights of individual training observations are truncated at 5.00%. Aggregation of IATEs to ATE and GATEs may not be exact due to weight truncation.

INFERENCE
Inference is based on using the weight matrix. Nonparametric regressions are based on k-nearest neighbours.

NOTE
Treatment effects for specific treatment groups (so-called treatment effects on the treated or non-treated) can only be provided if the data provided for prediction contains a treatment variable (which is not required for the other effects).

Section 2.2.2.1: ATE

Modified Causal Forest: Estimation

RESULT

ATE for sal_avg

Comparison	*Effect*	*SE*	*t-value*	*p-value (%)*	*Sig.*
1 vs 0	-7232.654	824.878	8.77	0.0	****
2 vs 0	-27914.232	1004.205	27.8	0.0	****
2 vs 1	-20681.578	1122.745	18.42	0.0	****

Note: *, **, ***, **** denote significance at the 10%, 5%, 1%, 0.1% level. The results for the potential outcomes can be found in the output files.

ATE for sal_3

Comparison	*Effect*	*SE*	*t-value*	*p-value (%)*	*Sig.*
1 vs 0	-27564.127	497.456	55.41	0.0	****
2 vs 0	-26672.186	541.816	49.23	0.0	****
2 vs 1	891.941	393.282	2.27	2.32	**

Note: *, **, ***, **** denote significance at the 10%, 5%, 1%, 0.1% level. The results for the potential outcomes can be found in the output files.

ATE for sal_4

Comparison	*Effect*	*SE*	*t-value*	*p-value (%)*	*Sig.*
1 vs 0	-31571.743	1086.896	29.05	0.0	****
2 vs 0	-44583.812	898.48	49.62	0.0	****
2 vs 1	-13012.07	1037.891	12.54	0.0	****

Note: *, **, ***, **** denote significance at the 10%, 5%, 1%, 0.1% level. The results for the potential outcomes can be found in the output files.

ATE for sal_5

Comparison	*Effect*	*SE*	*t-value*	*p-value (%)*	*Sig.*
1 vs 0	-7732.474	1217.346	6.35	0.0	****
2 vs 0	-44817.87	970.928	46.16	0.0	****
2 vs 1	-37085.396	1329.913	27.89	0.0	****

Note: *, **, ***, **** denote significance at the 10%, 5%, 1%, 0.1% level. The results for the potential outcomes can be found in the output files.

ATE for sal_6

Comparison	*Effect*	*SE*	*t-value*	*p-value (%)*	*Sig.*

1 vs 0	-813.187	1185.08	0.69	49.02	
2 vs 0	-39476.997	1295.088	30.48	0.0	****
2 vs 1	-38663.81	1590.445	24.31	0.0	****

Note: *, **, ***, **** denote significance at the 10%, 5%, 1%, 0.1% level. The results for the potential outcomes can be found in the output files.

ATE for sal_7

Comparison	*Effect*	*SE*	*t-value*	*p-value (%)*	*Sig.*
1 vs 0	2406.494	1140.867	2.11	3.49	**
2 vs 0	-27550.602	1775.65	15.52	0.0	****
2 vs 1	-29957.096	2004.068	14.95	0.0	****

Note: *, **, ***, **** denote significance at the 10%, 5%, 1%, 0.1% level. The results for the potential outcomes can be found in the output files.

ATE for sal_8

Comparison	*Effect*	*SE*	*t-value*	*p-value (%)*	*Sig.*
1 vs 0	5662.066	1137.998	4.98	0.0	****
2 vs 0	-15072.196	2025.062	7.44	0.0	****
2 vs 1	-20734.261	2256.476	9.19	0.0	****

Note: *, **, ***, **** denote significance at the 10%, 5%, 1%, 0.1% level. The results for the potential outcomes can be found in the output files.

ATE for sal_9

Comparison	*Effect*	*SE*	*t-value*	*p-value (%)*	*Sig.*
1 vs 0	8945.505	1175.191	7.61	0.0	****
2 vs 0	2922.404	1927.8	1.52	12.85	
2 vs 1	-6023.101	2211.256	2.72	0.65	***

Note: *, **, ***, **** denote significance at the 10%, 5%, 1%, 0.1% level. The results for the potential outcomes can be found in the output files.

ATE for empl_ttl

Comparison	*Effect*	*SE*	*t-value*	*p-value (%)*	*Sig.*
1 vs 0	0.432	0.376	1.15	25.01	
2 vs 0	-10.274	0.463	22.19	0.0	****
2 vs 1	-10.706	0.468	22.88	0.0	****

Note: *, **, ***, **** denote significance at the 10%, 5%, 1%, 0.1% level. The results for the potential outcomes can be found in the output files.

Modified Causal Forest: Estimation

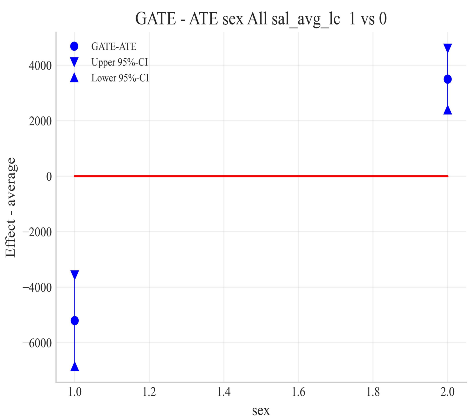
ATE for empl_chge

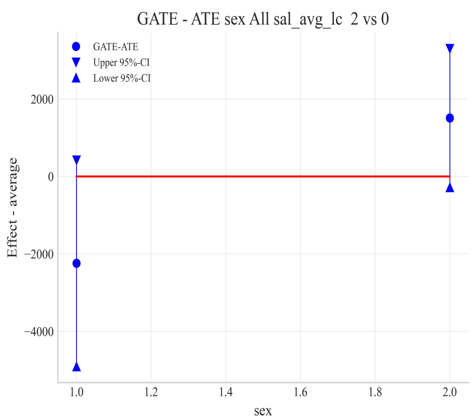
Comparison	*Effect*	*SE*	*t-value*	*p-value (%)*	*Sig.*
1 vs 0	0.15	0.171	0.88	37.89	
2 vs 0	-1.685	0.156	10.8	0.0	****
2 vs 1	-1.835	0.203	9.04	0.0	****

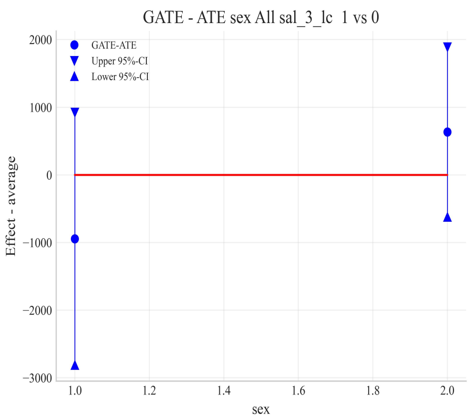
Note: *, **, ***, **** denote significance at the 10%, 5%, 1%, 0.1% level. The results for the potential outcomes can be found in the output files.

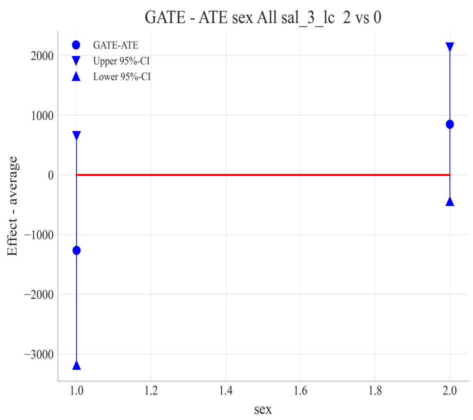
Section 2.2.2.2: GATE

GATE

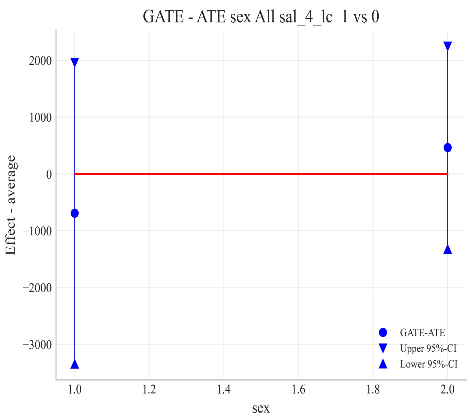


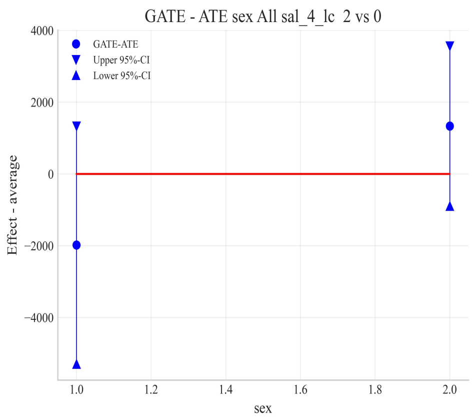


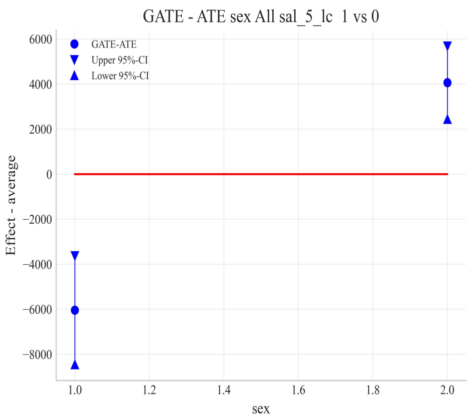


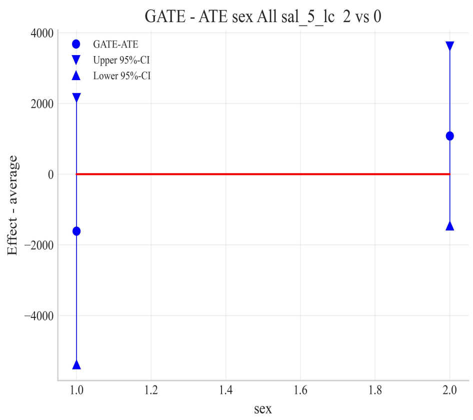


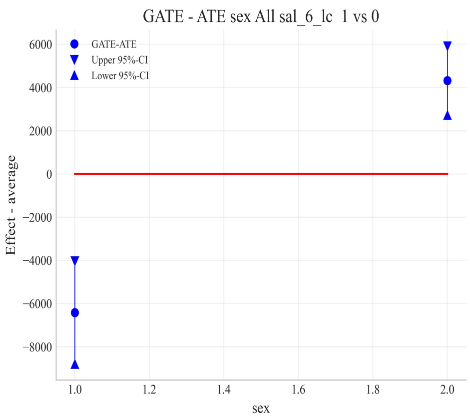
Modified Causal Forest: Estimation

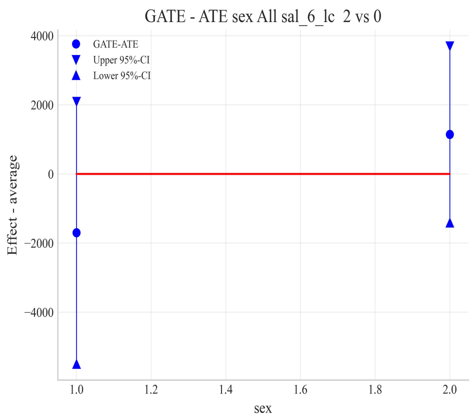


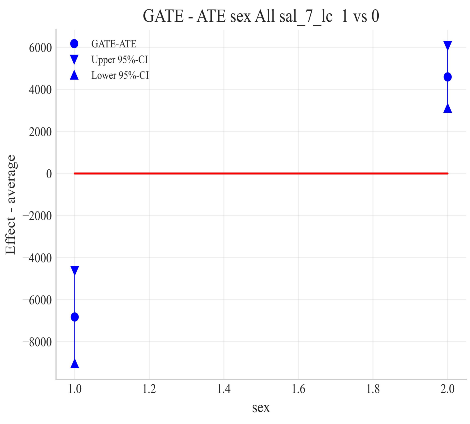


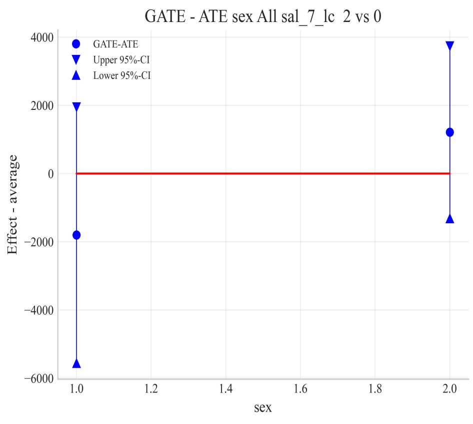




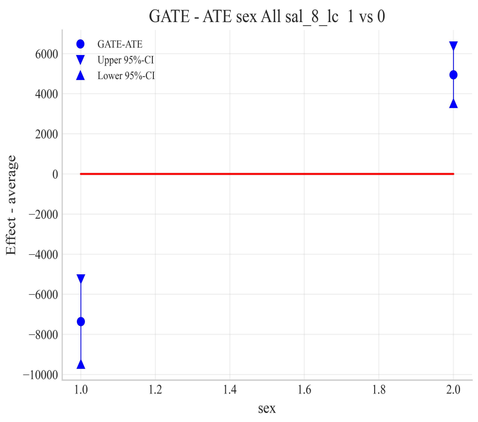


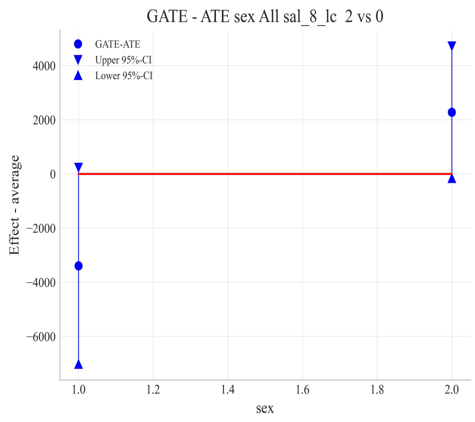


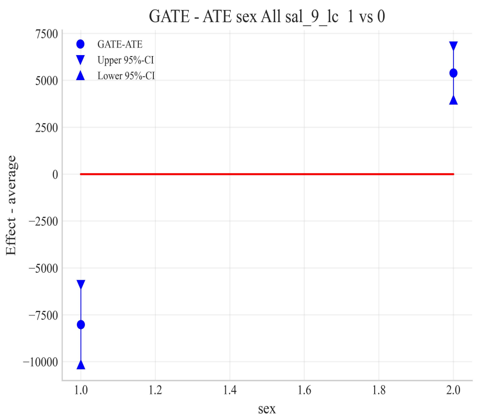


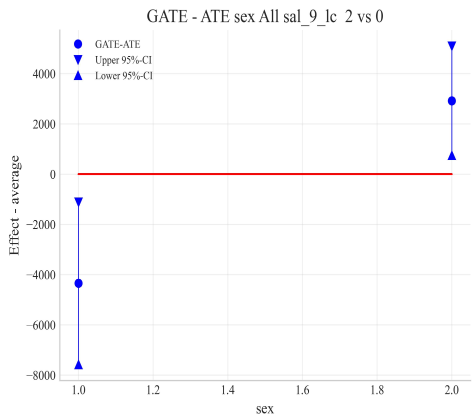


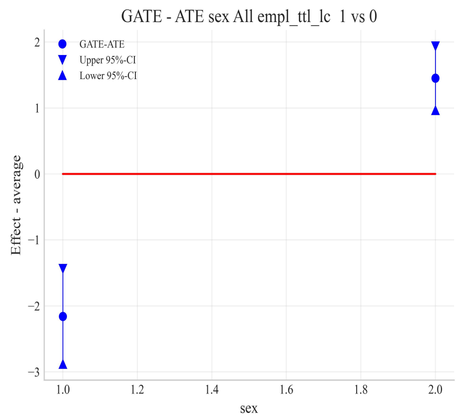
Modified Causal Forest: Estimation

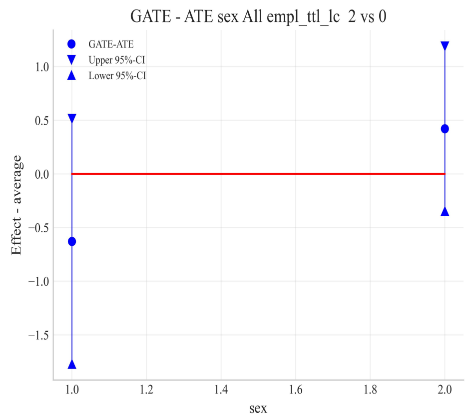


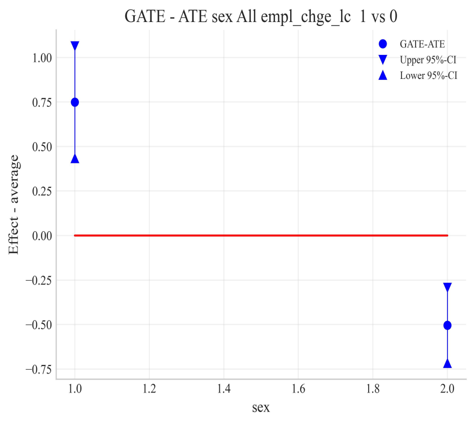


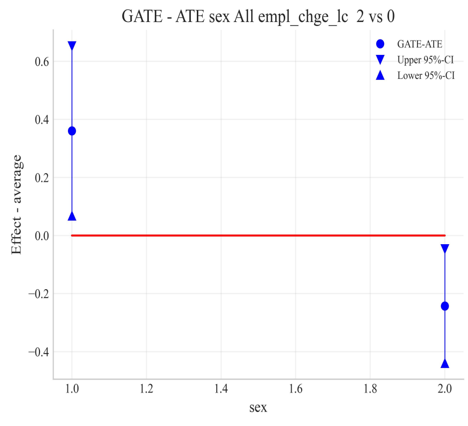




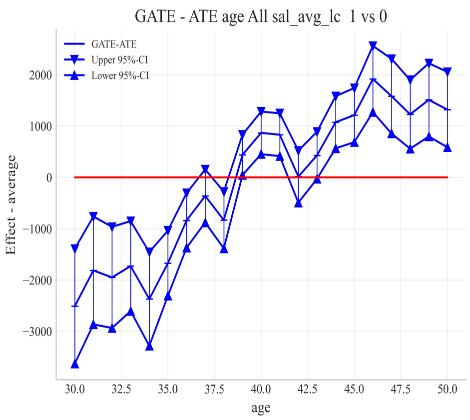


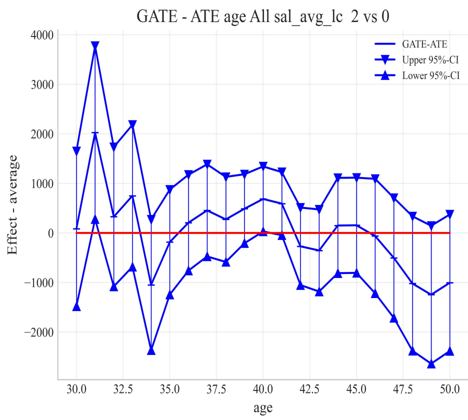


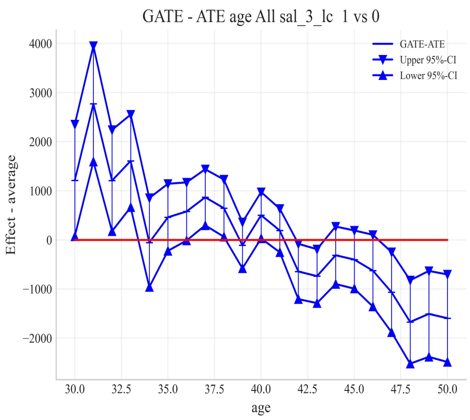


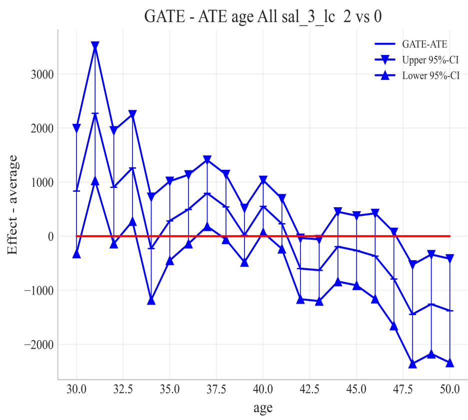


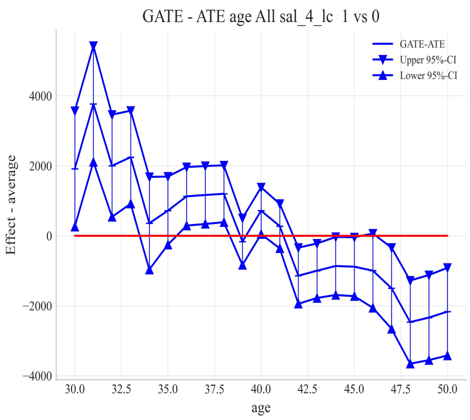
Modified Causal Forest: Estimation

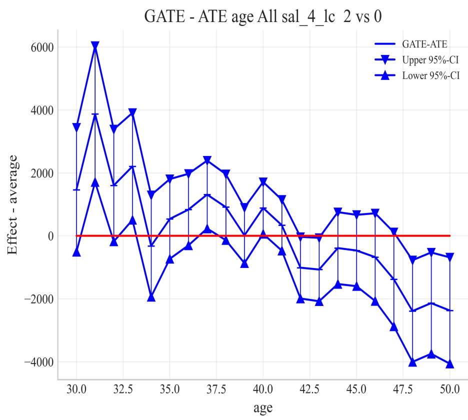


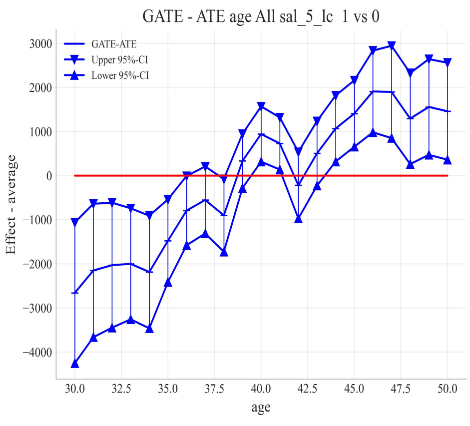


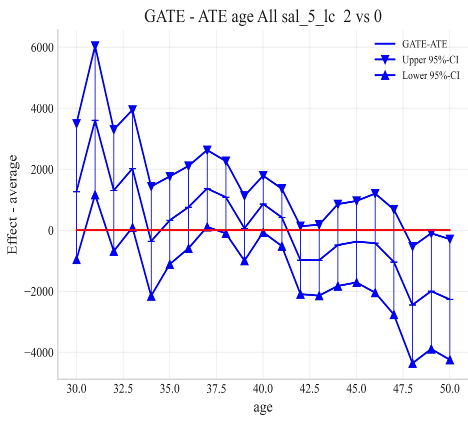




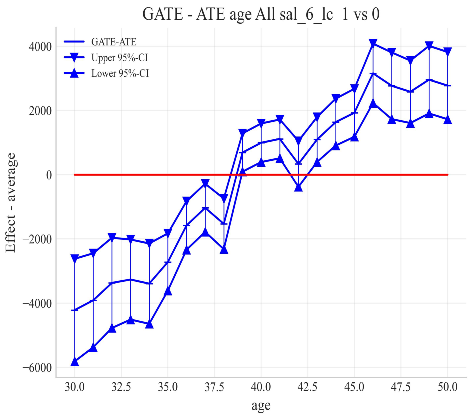


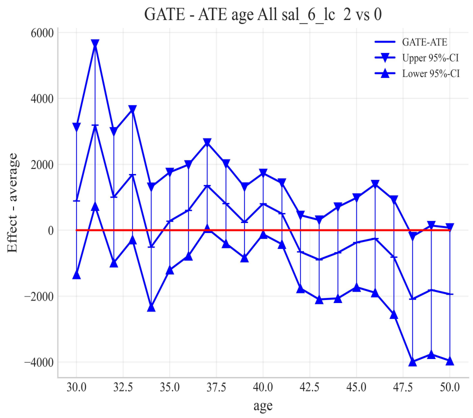


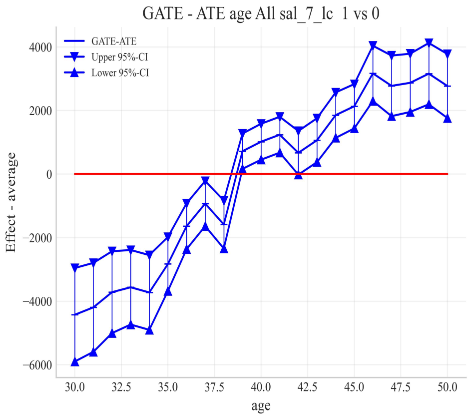


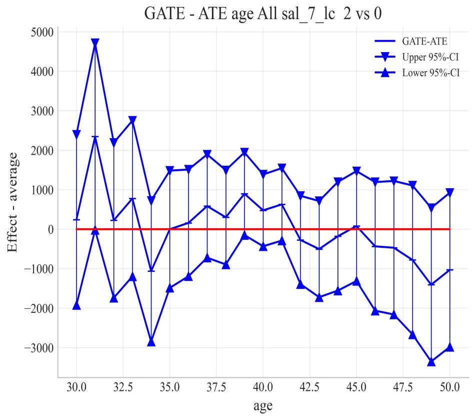


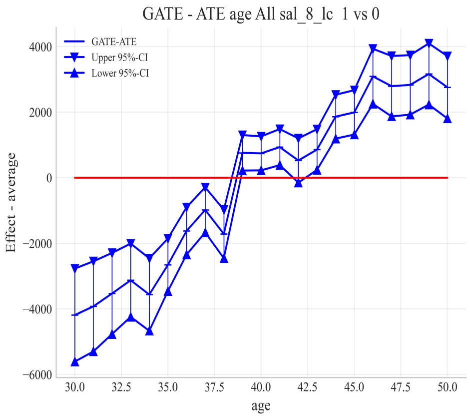
Modified Causal Forest: Estimation

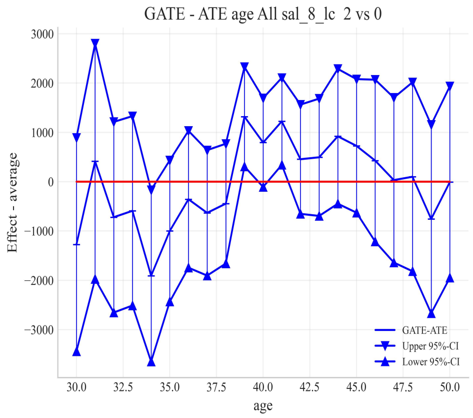


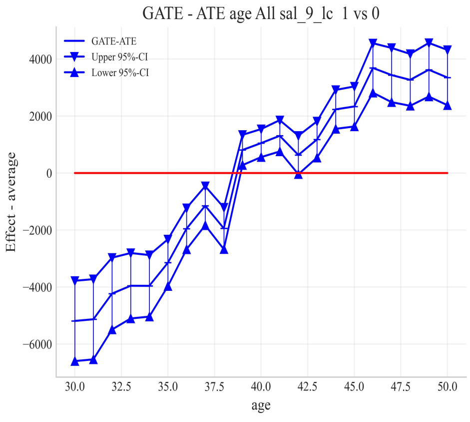


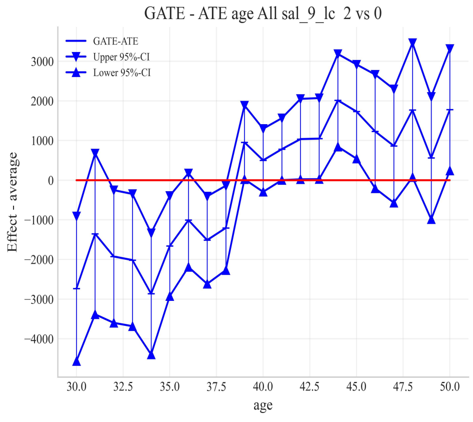




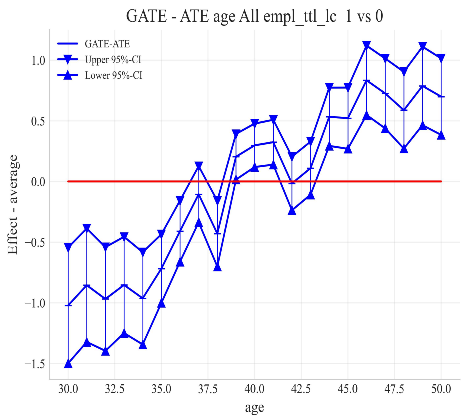


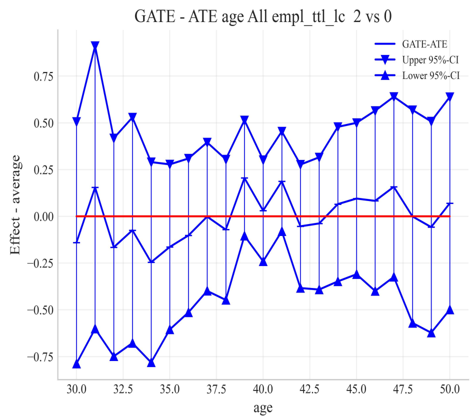


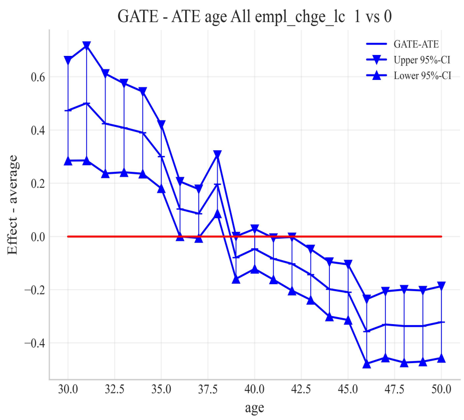


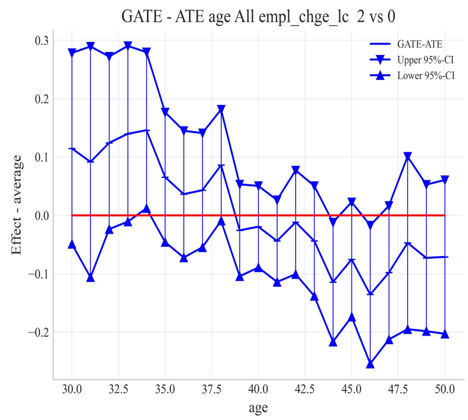


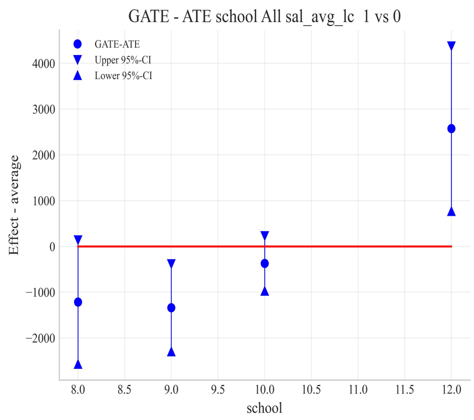
Modified Causal Forest: Estimation

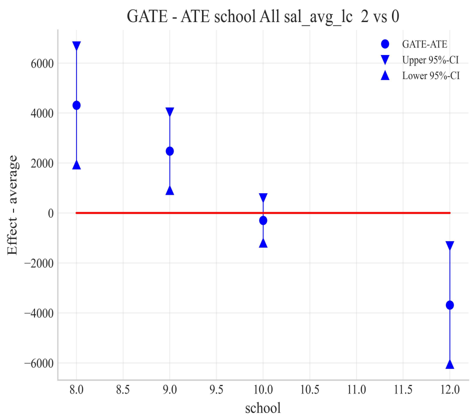


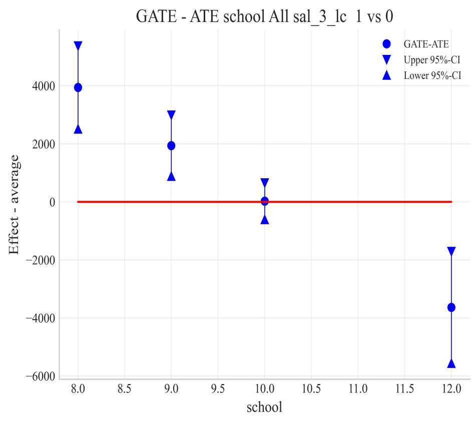


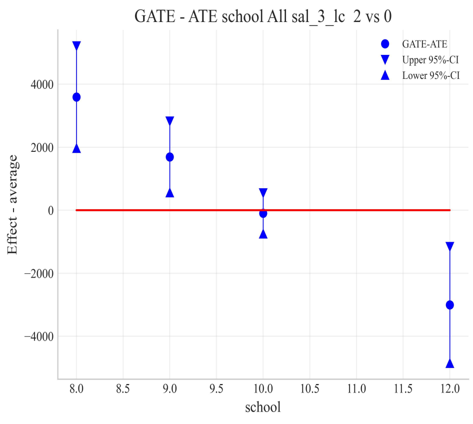




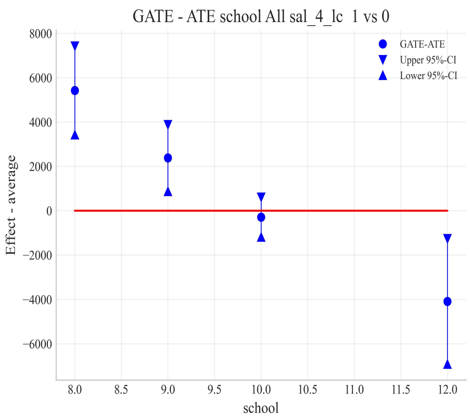


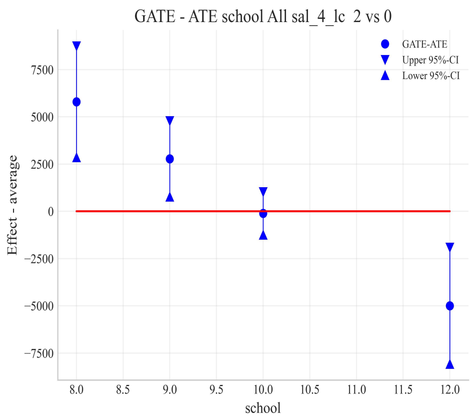


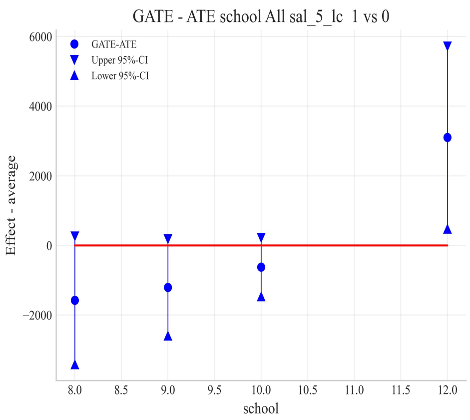


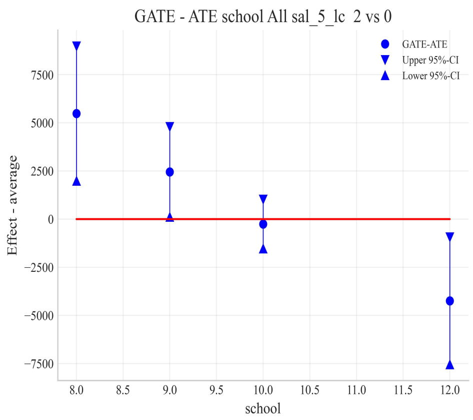


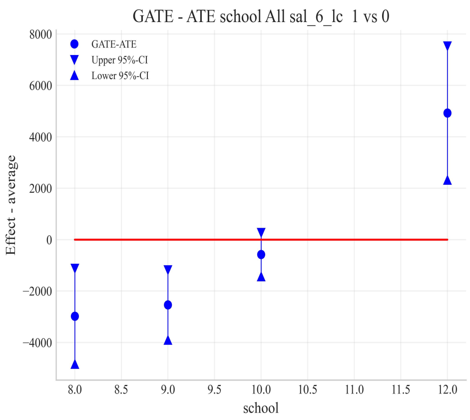
Modified Causal Forest: Estimation

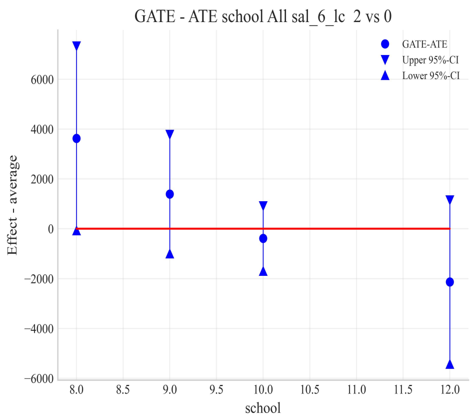


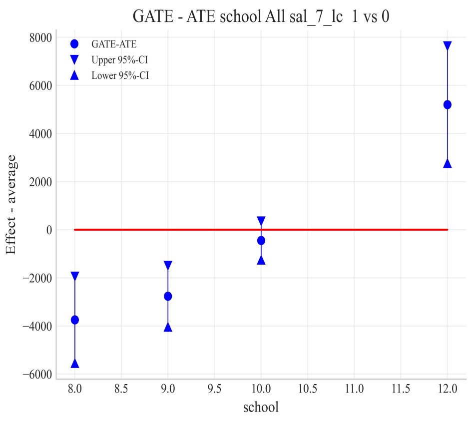


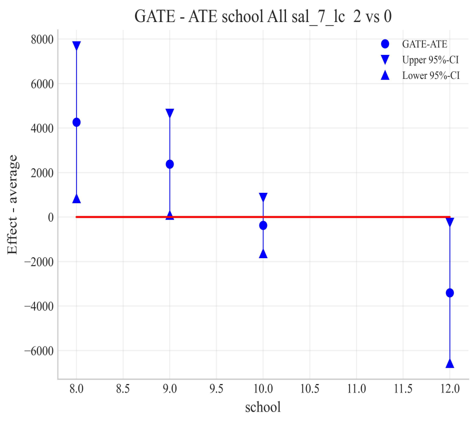




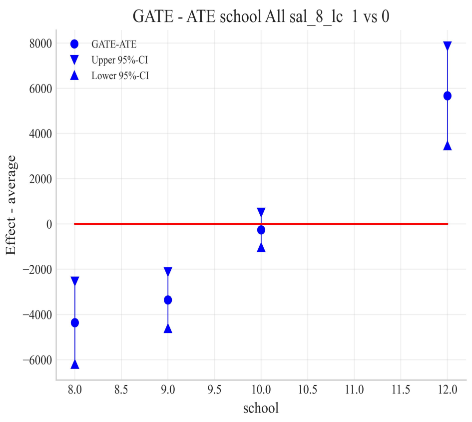


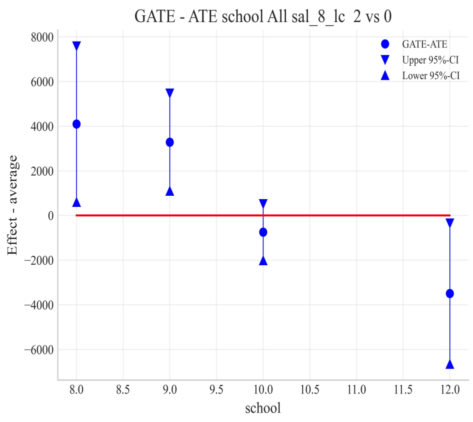


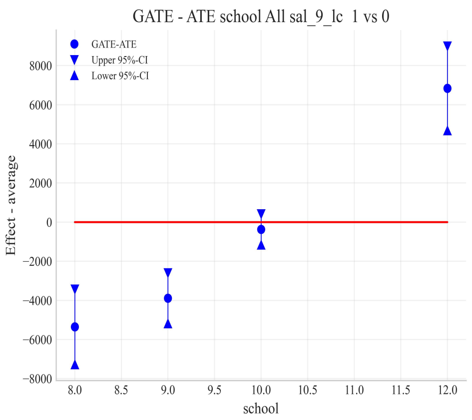


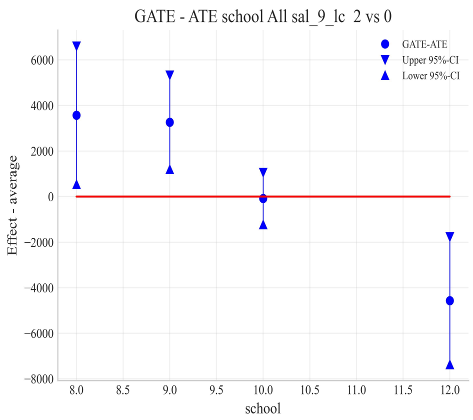


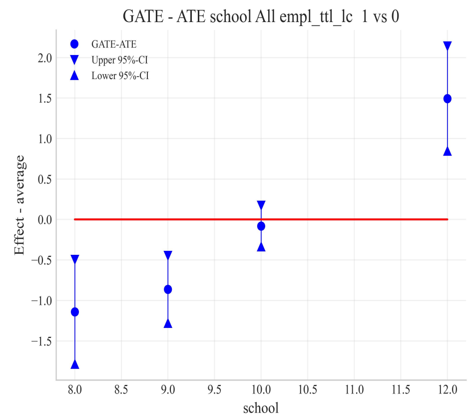
Modified Causal Forest: Estimation

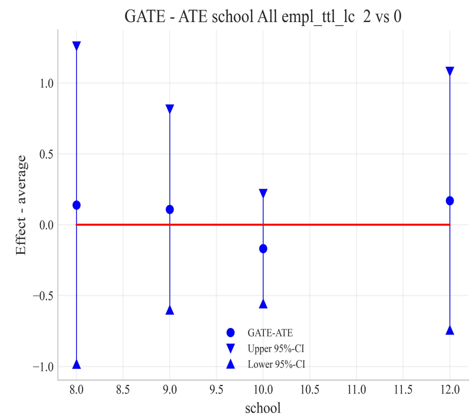


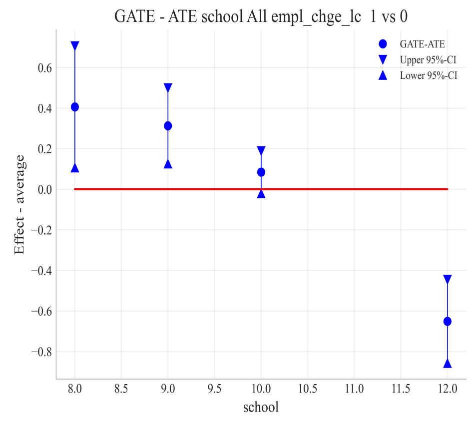


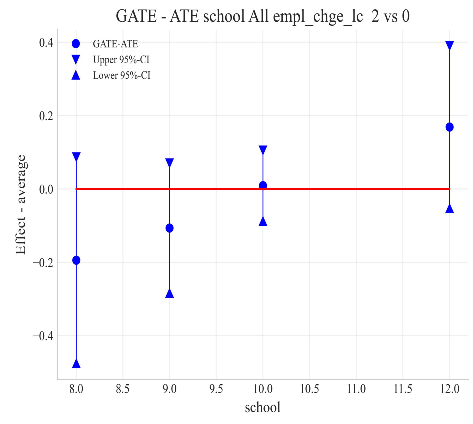




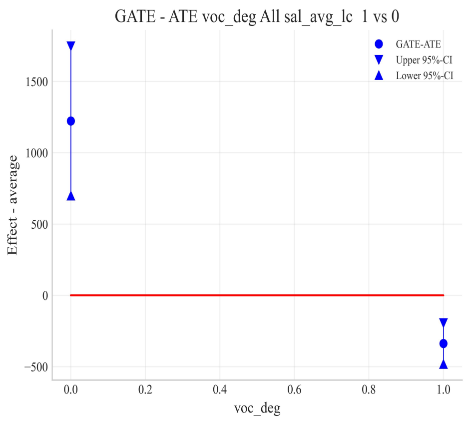


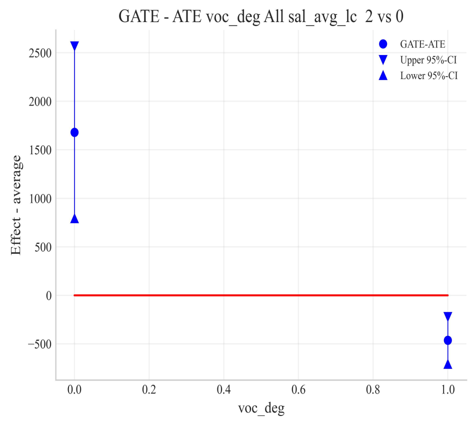


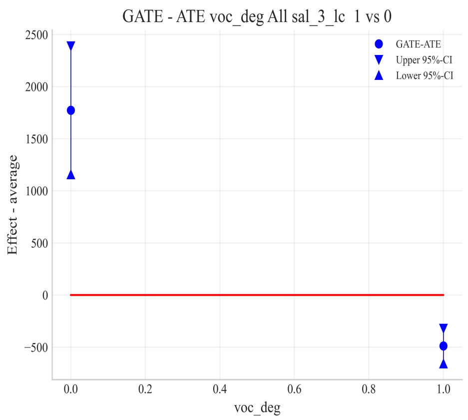


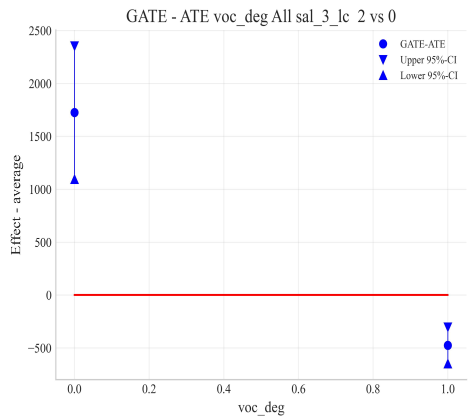


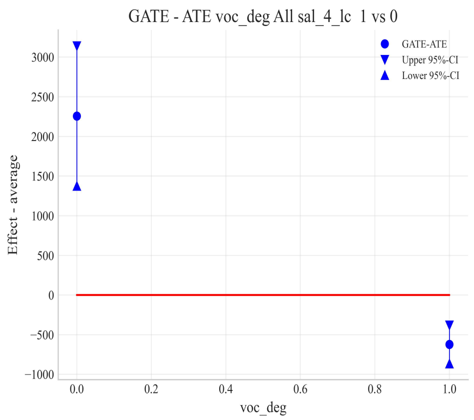
Modified Causal Forest: Estimation

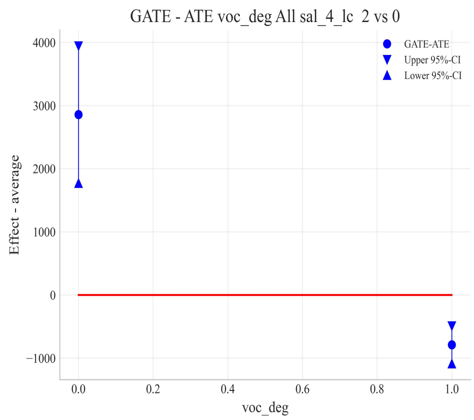


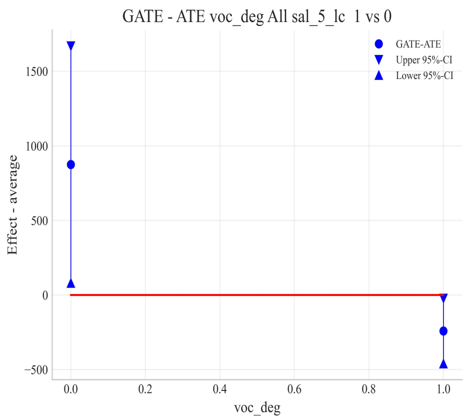


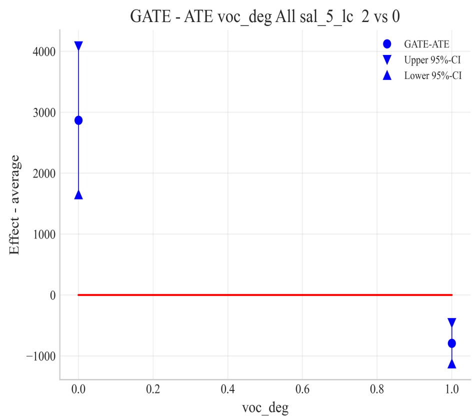




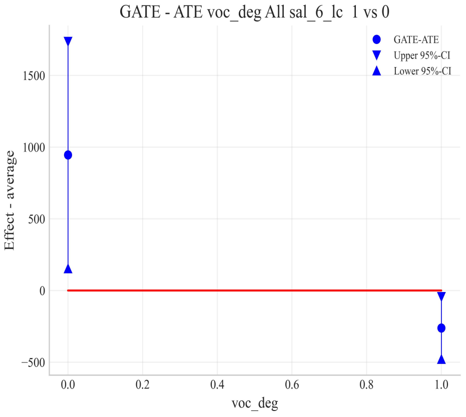


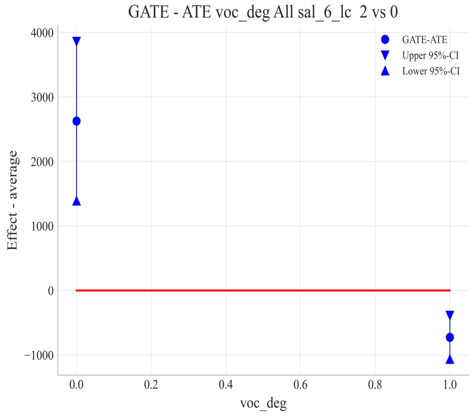


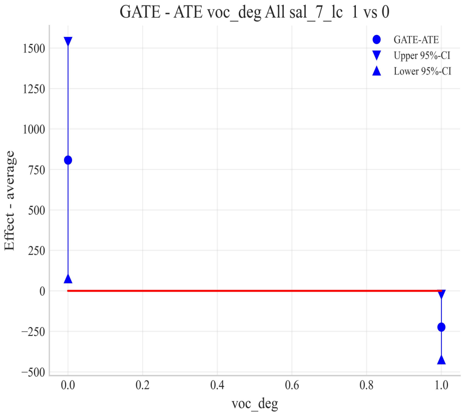


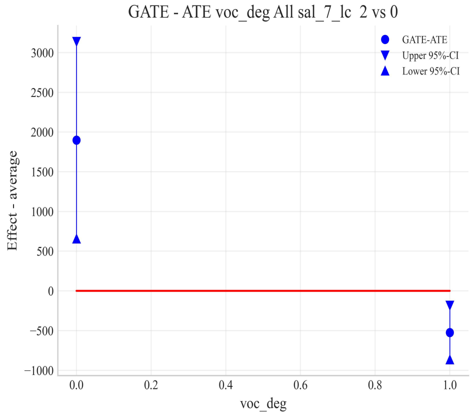


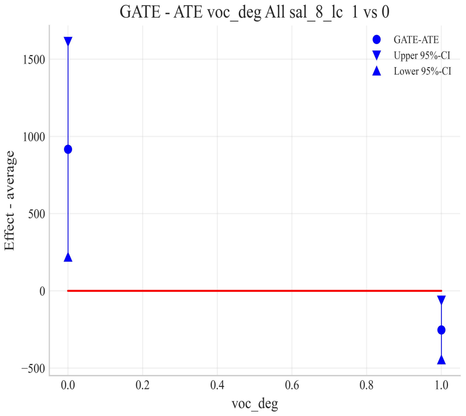
Modified Causal Forest: Estimation

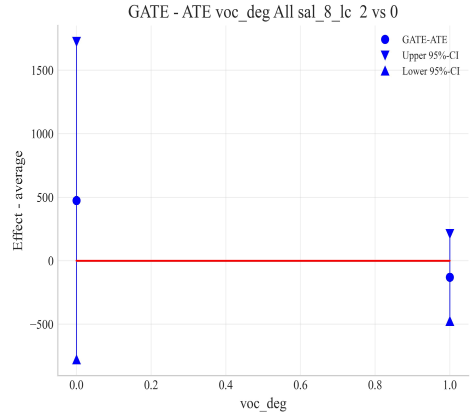


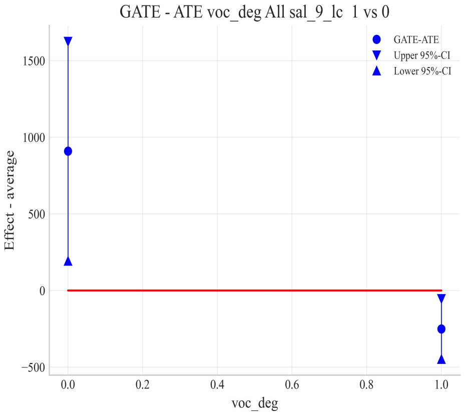


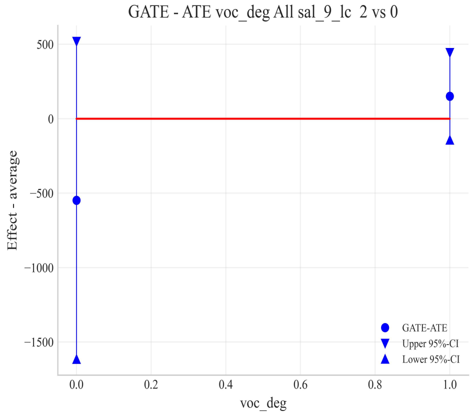




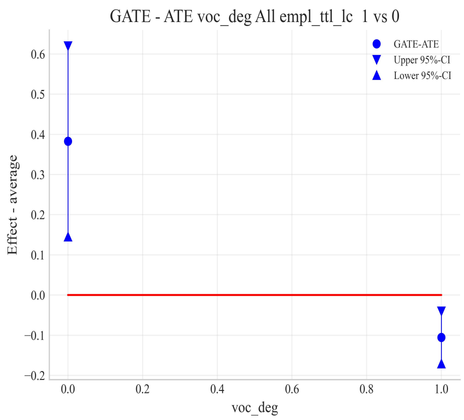


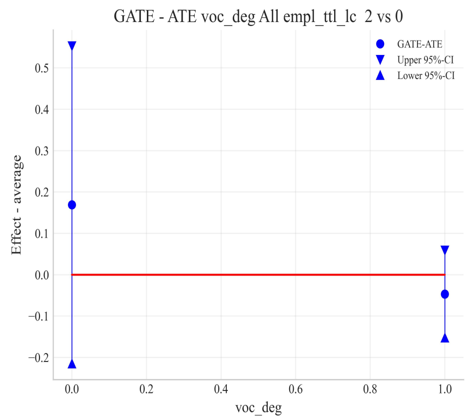


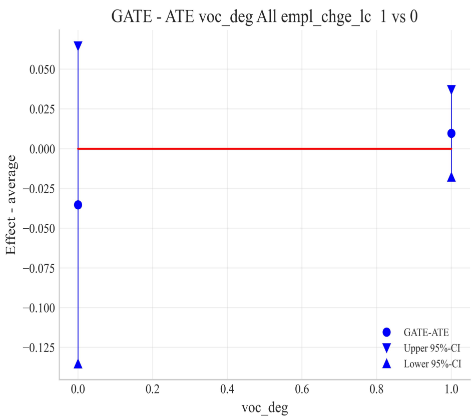


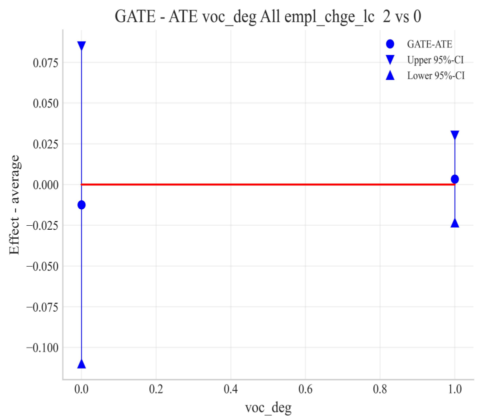


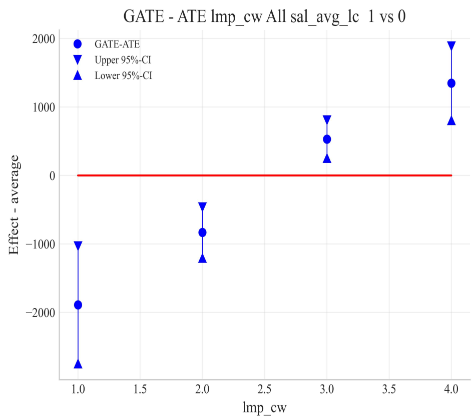
Modified Causal Forest: Estimation

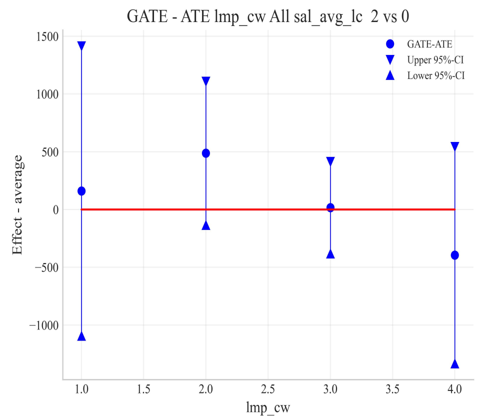


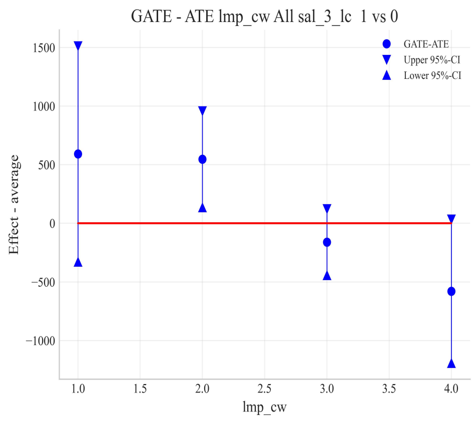


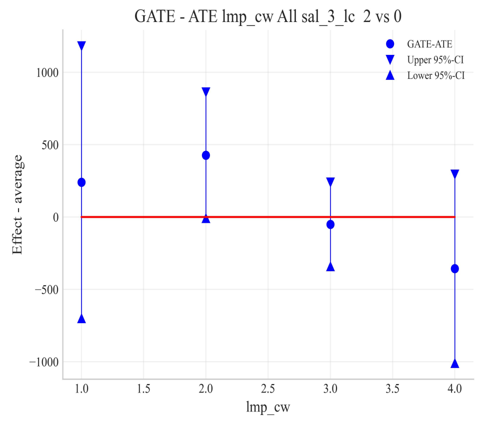




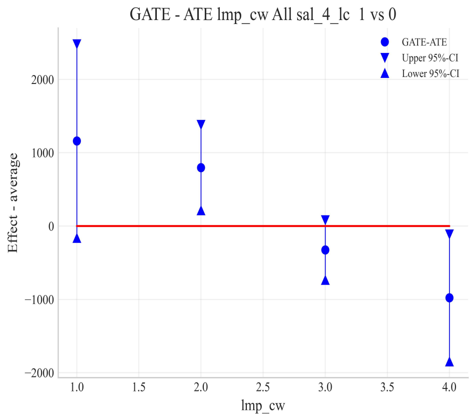


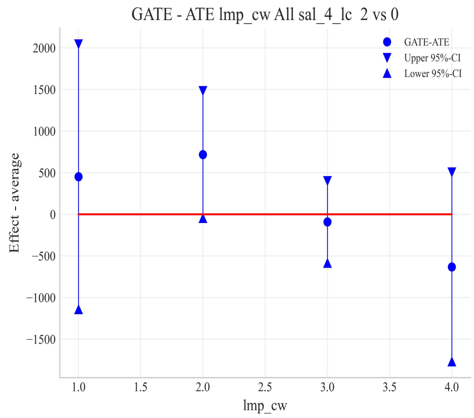


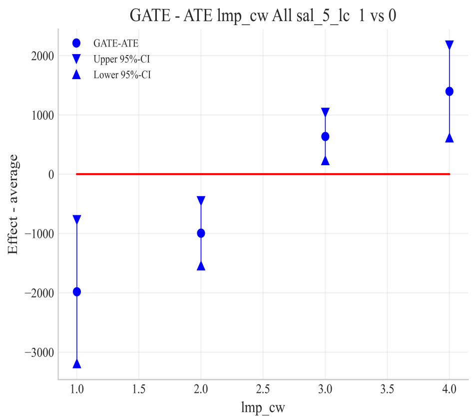


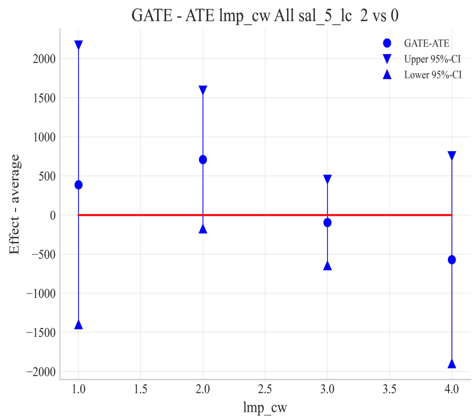


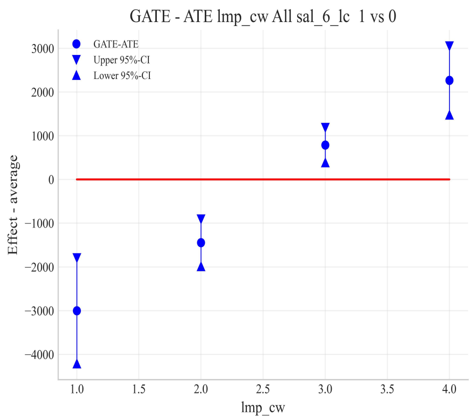
Modified Causal Forest: Estimation

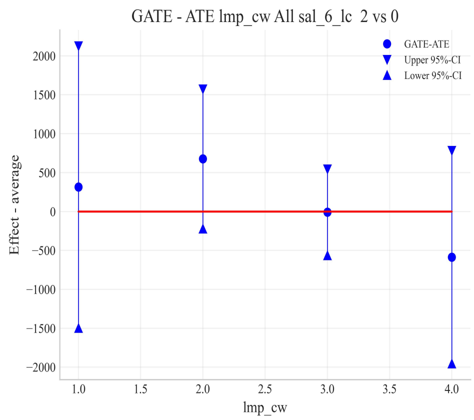


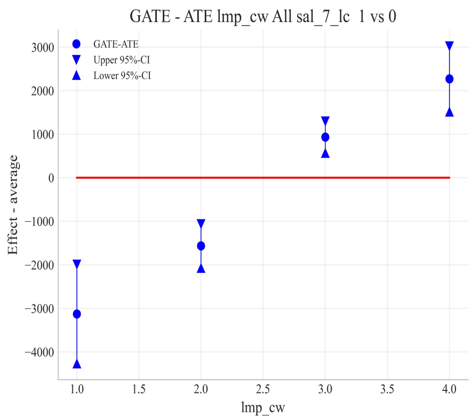


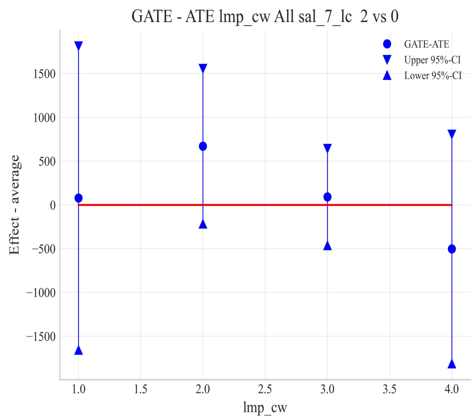




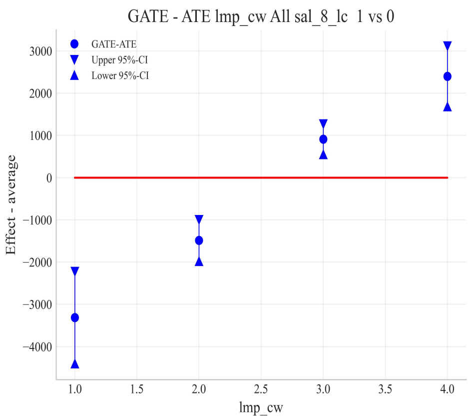


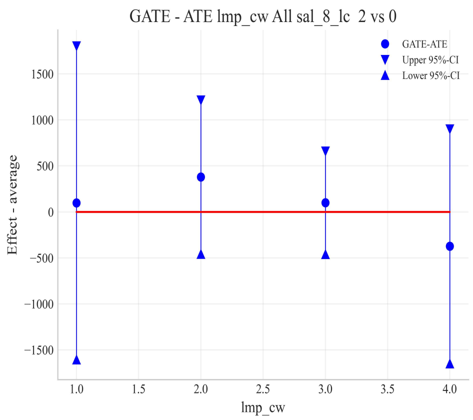


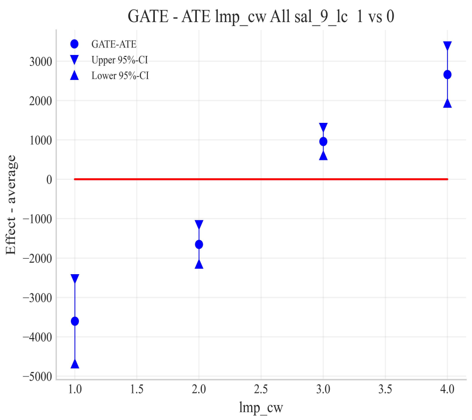


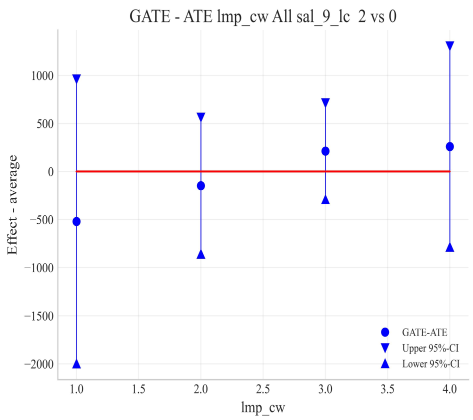


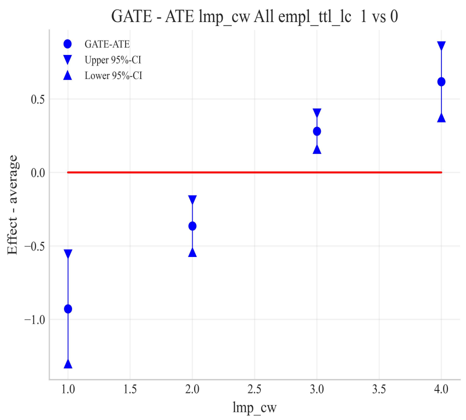
Modified Causal Forest: Estimation

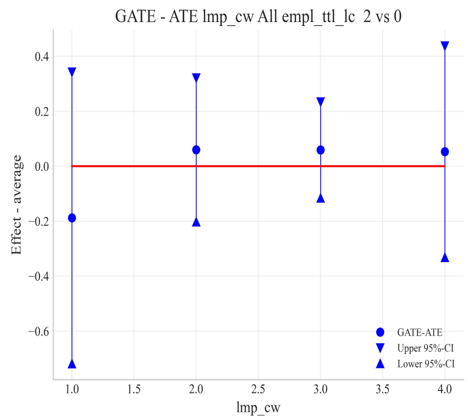


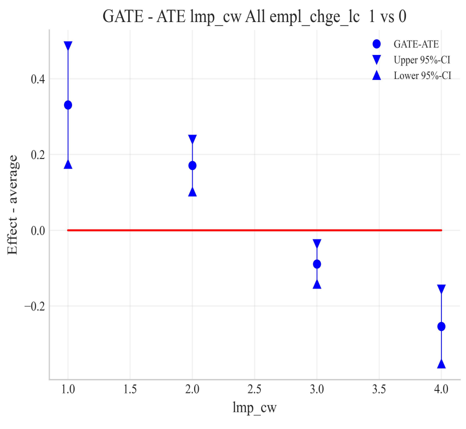


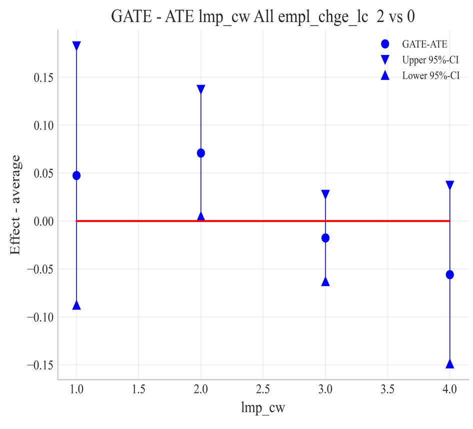




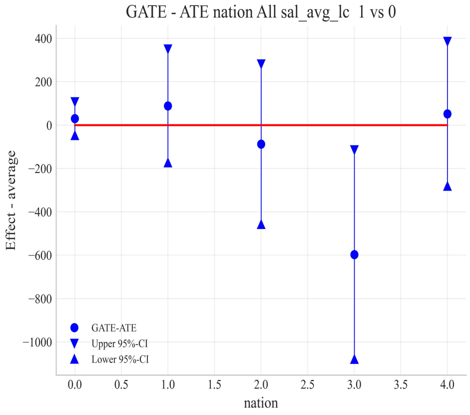


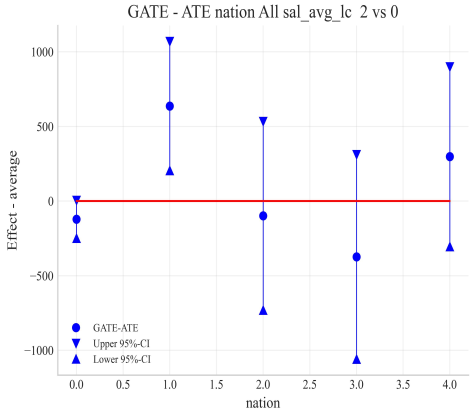


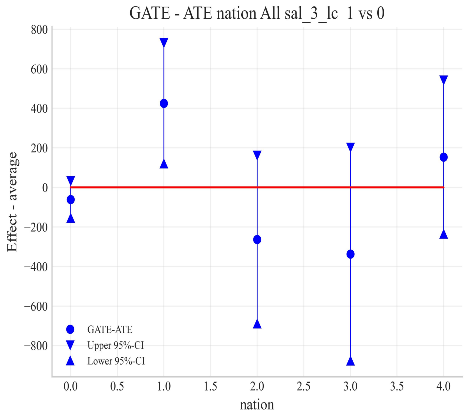


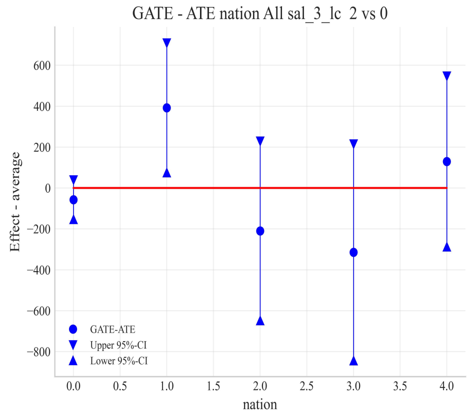


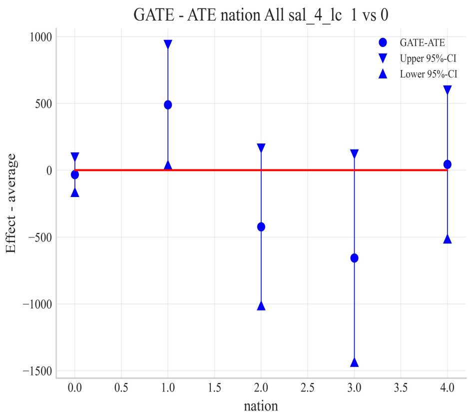
Modified Causal Forest: Estimation

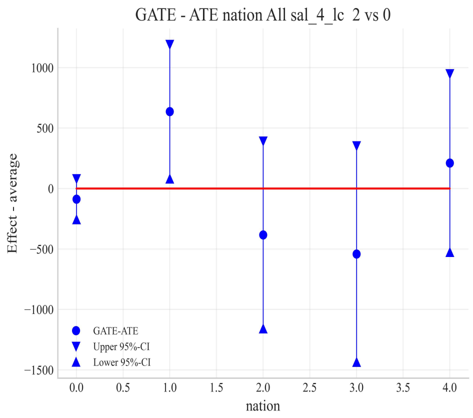


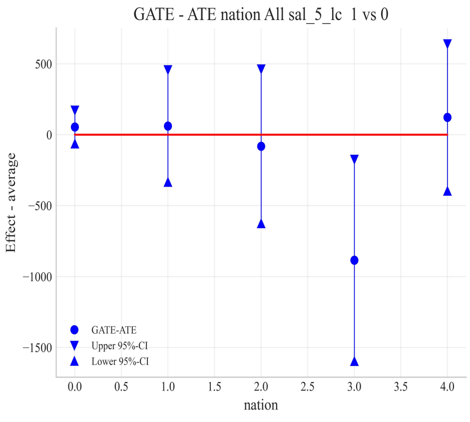


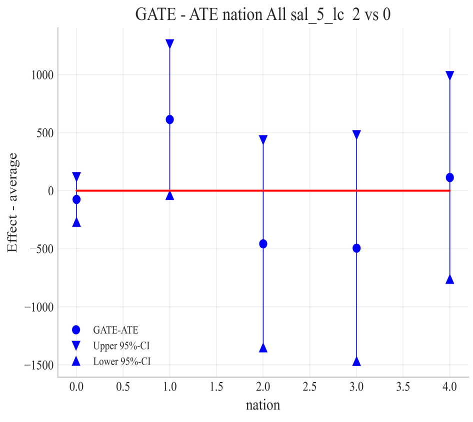




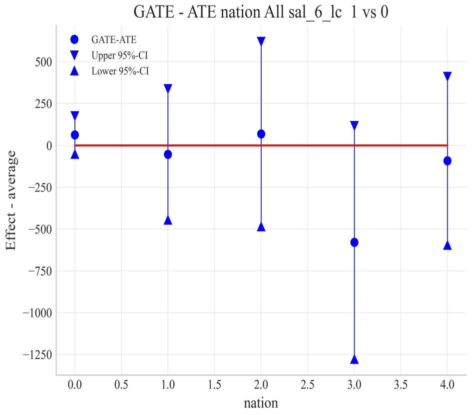


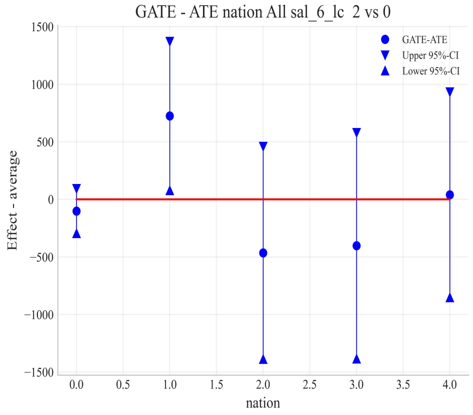


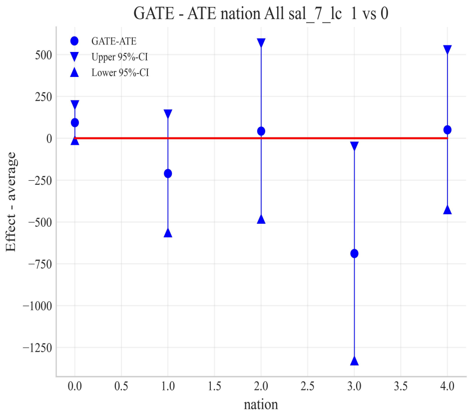


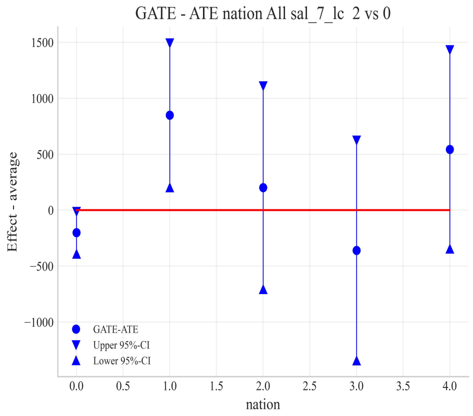


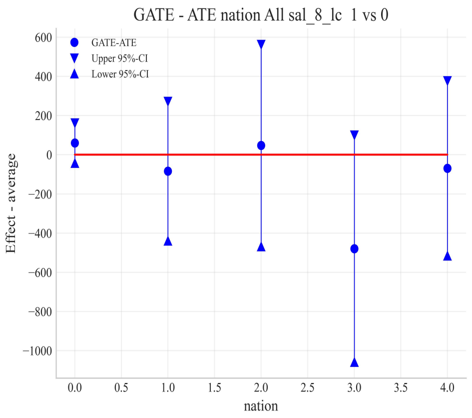
Modified Causal Forest: Estimation

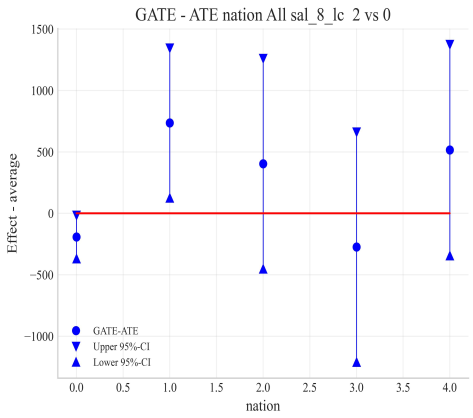


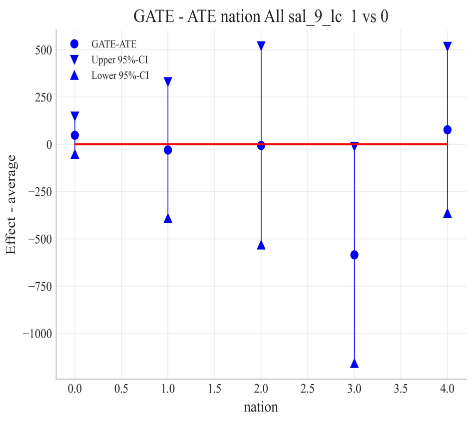


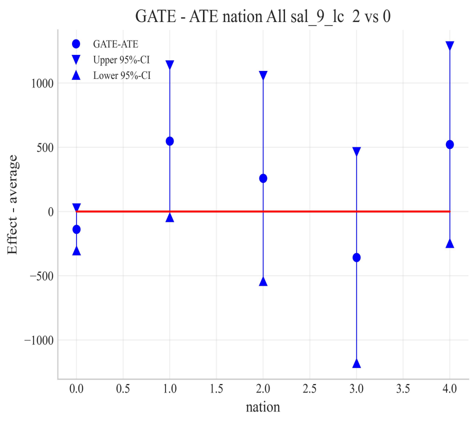




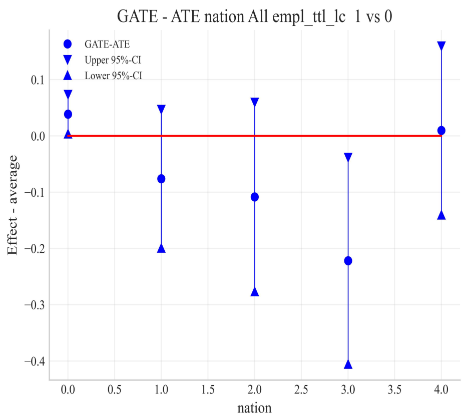


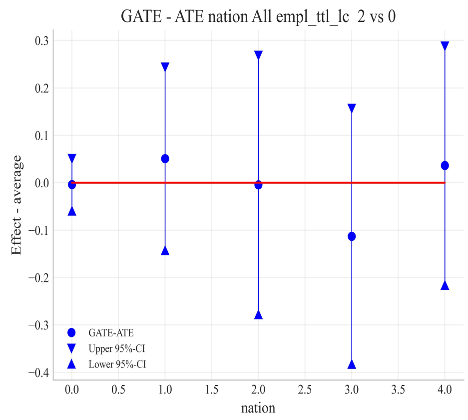


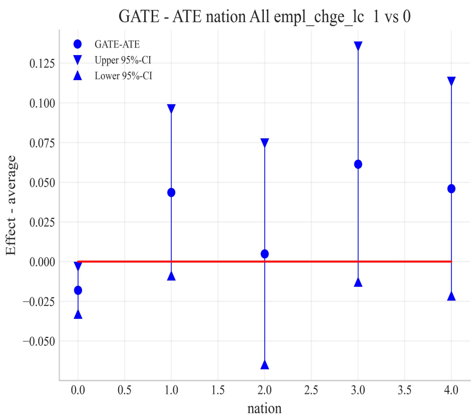


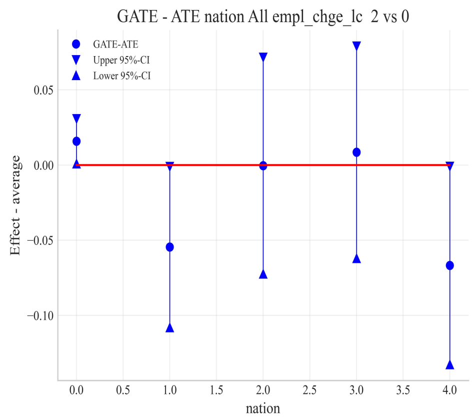


Modified Causal Forest: Estimation









Note: Detailed tables and figures for additional effects are contained in the output files and output directories.

Section 2.2.2.3: IATE

This section contains parts of the descriptive analysis of the IATEs. Use the analyse method to obtain more descriptives of the IATEs, like their distribution, and their relations to the features.

RESULTS

Outcome variable: sal_avg

Comparison	Mean	Median	Std	Effect > 0	mean(SE)	sig 10%	sig 5%	sig 1%
1 vs 0	-7220.94581	-5618.73976	5173.46349	1.50%	1958.63909	82.02%	77.26%	64.74%
2 vs 0	-27892.98229	-27301.24137	6895.36558	0.00%	3139.31972	100.00%	100.00%	100.00%
2 vs 1	-20672.03648	-19904.39476	6421.24052	0.00%	3234.60759	99.65%	99.12%	97.44%

Outcome variable: sal_3

Modified Causal Forest: Estimation

Comparison	Mean	Median	Std	Effect > 0	mean(SE)	sig 10%	sig 5%	sig 1%
1 vs 0	-27546.43268	-26466.56055	6613.74004	0.00%	2140.64149	100.00%	100.00%	100.00%
2 vs 0	-26654.43333	-25560.53929	6172.26973	0.00%	2198.60058	100.00%	100.00%	100.00%
2 vs 1	891.99935	868.74373	706.81620	90.79%	2088.89919	0.00%	0.00%	0.00%

Outcome variable: sal_4

Comparison	Mean	Median	Std	Effect > 0	mean(SE)	sig 10%	sig 5%	sig 1%
1 vs 0	-31562.01146	-29724.36703	8299.79374	0.00%	3182.12375	100.00%	100.00%	100.00%
2 vs 0	-44553.92498	-42710.22468	10257.68103	0.00%	3790.51035	100.00%	100.00%	100.00%
2 vs 1	-12991.91352	-13090.75625	2661.33849	0.00%	3711.27241	99.82%	97.58%	87.88%

Outcome variable: sal_5

Comparison	Mean	Median	Std	Effect > 0	mean(SE)	sig 10%	sig 5%	sig 1%
1 vs 0	-7722.67696	-5391.63146	5731.98373	1.28%	2972.26331	67.03%	58.53%	44.07%
2 vs 0	-44824.76400	-43405.73744	10007.44970	0.00%	4272.88036	100.00%	100.00%	100.00%
2 vs 1	-37102.08704	-36944.25243	10747.75134	0.00%	4631.96986	100.00%	100.00%	100.00%

Outcome variable: sal_6

Comparison	Mean	Median	Std	Effect > 0	mean(SE)	sig 10%	sig 5%	sig 1%
1 vs 0	-803.99679	1285.86905	6619.97823	56.10%	2872.44973	52.71%	41.47%	28.07%
2 vs 0	-39441.61910	-39341.01009	8269.11063	0.00%	4416.04338	100.00%	100.00%	100.00%
2 vs 1	-38637.62230	-38475.78177	10148.20325	0.00%	4894.61990	100.00%	100.00%	100.00%

Outcome variable: sal_7

Comparison	Mean	Median	Std	Effect > 0	mean(SE)	sig 10%	sig 5%	sig 1%
1 vs 0	2416.55265	4529.42873	6950.35527	67.92%	2710.36465	70.16%	64.65%	51.39%
2 vs 0	-27515.86064	-27452.25735	6836.14425	0.00%	4574.47793	100.00%	100.00%	100.00%
2 vs 1	-29932.41329	-29295.38840	9763.19130	0.00%	5086.08208	98.50%	97.62%	95.99%

Outcome variable: sal_8

Modified Causal Forest: Estimation

Comparison	Mean	Median	Std	Effect > 0	mean(SE)	sig 10%	sig 5%	sig 1%
1 vs 0	5677.18086	7672.50925	7432.21066	75.98%	2657.85517	74.79%	70.38%	60.73%
2 vs 0	-15012.56376	-13779.08054	6014.07618	0.00%	4684.86308	93.79%	85.81%	65.98%
2 vs 1	-20689.74462	-20360.05822	8354.34863	0.88%	5264.90234	92.46%	90.57%	84.53%

Outcome variable: sal_9

Comparison	Mean	Median	Std	Effect > 0	mean(SE)	sig 10%	sig 5%	sig 1%
1 vs 0	8958.54666	10747.31152	8317.57165	82.42%	2716.78924	77.79%	73.64%	66.64%
2 vs 0	2924.59662	3729.97444	5664.07604	70.60%	4181.21061	34.24%	24.33%	7.71%
2 vs 1	-6033.95004	-5011.19602	8412.27511	21.68%	4900.17187	40.86%	34.11%	22.26%

Outcome variable: empl_ttl

Comparison	Mean	Median	Std	Effect > 0	mean(SE)	sig 10%	sig 5%	sig 1%
1 vs 0	0.44306	1.07409	2.16254	65.62%	0.92975	62.32%	53.50%	35.70%
2 vs 0	-10.25718	-10.13176	0.88776	0.00%	1.36279	100.00%	100.00%	100.00%
2 vs 1	-10.70024	-11.14366	1.72019	0.00%	1.37976	100.00%	100.00%	100.00%

Outcome variable: empl_chge

Comparison	Mean	Median	Std	Effect > 0	mean(SE)	sig 10%	sig 5%	sig 1%
1 vs 0	0.14799	-0.04291	0.82355	47.77%	0.39358	44.73%	37.68%	24.46%
2 vs 0	-1.68100	-1.70495	0.40529	0.00%	0.37093	99.52%	98.81%	94.89%
2 vs 1	-1.82900	-1.84704	0.73164	0.00%	0.47079	93.79%	90.92%	85.19%