Machine Learning in Mediation Analysis: An Extension of Inverse Probability Weighting

DSE4231 Topics in Data Science and the Digital Economy

AY 2024/2025 Semester 2

14 April 2025

Eliza Ong, A0238948J

Krystal Low Li Tong, A0238015M

Lily Rozana Joehann Aung, A0239463X

Yang Shu Ting, A0238974L

1. Introduction

1.1 Motivation

A mediator in causal inference is an intermediate variable between a treatment and the outcome, it may also be regarded as an intermediate outcome (Robins and Greenland, 1992). However, random treatment assignment does not also mean that the mediator is random. This presents a challenge in identifying the total treatment effect by conditioning on a mediator due to confounding of the mediator and outcome (Rosenbaum, 1984). Most studies on this topic however, depend on inflexible models with strong linear specifications imposed (Pearl, 2001). In Huber's (2014) paper: 'Identifying causal mechanisms (primarily) based on inverse probability weighting', inverse probability weighting (IPW) was leveraged to disentangle the direct effect of the treatment on the outcome and the indirect ones that run through discrete or continuous mediators in experiments. With that, the extent of the total effect contributed by mediators can be found without linearity assumptions. This evaluation of direct and indirect effects is frequently referred to as mediation analysis.

2. Literature Review

His paper has several key contributions. First, identification results are easily obtained through IPW by semi parametric or nonparametric propensity scores. Secondly, a functional form restriction is presented to allow for the identification of indirect effects with the assumption that any confounders affecting the mediator and the outcome can be accounted for by controlling for variables measured before the treatment is assigned. Third, IPW is proven to still identify a partial indirect effect when confounders are controlled for (i.e., the portion of indirect effect that is uncorrelated to confounders). Lastly, a simulation study which evaluates the direct and indirect effects of the Job Corps program is included as an empirical contribution.

To begin, identification relies on two main assumptions. The first is that the mediator is conditionally exogenous given the observed covariates. This means that the mediator and potential outcomes for a particular mediator state are conditionally independent given the covariates and the treatment. Next, Identification of indirect effects depends on whether the covariates are a function of the treatment.If they are a function of the treatment, the identification of the "total" indirect effect requires additional restrictions as correlations between the covariates and mediator are accounted for (Avin, Shpitser, and Pearl, 2005). If not, the "partial" indirect effect, which only considers the immediate link between the treatment and the mediator, is identified under weaker assumptions.

2.1 Methodology

In IPW, units are weighed by the inverse of their propensity score to be observed in a specific treatment state, given the mediator and the observed covariates. Our parameters of interest are the average treatment effect (ATE) on a binary treatment indicator D on some outcome variable Y. The goal is to disentangle ATE into a direct indirect effect operating through the mediator M which may be discrete or continuous. To define these parameters, the potential framework outcome by Rubin (1974) is utilised: Y(d), M(d) denote the potential outcome and the potential mediator state under treatment d $\in$ {0,1} respectively. For each unit, only one potential outcome and mediator state is observed. Additionally, the authors account for cases whereby the treatment changes characteristics which might affect the outcome. As such, the mediator and outcome is written as a function of X. Now, the total indirect effect comprises of all effects via M which either come from D or through X, and the partial indirect effect only identifies the effect from D going through M.

The realised outcome values, mediator values, ATE, total and partial indirect effect are written as such:

1. **Realised outcome values:** $Y(d, M(d)) = Y(d, M(d, X(d)), X(d))$
2. **Realised mediator values:** $M(d) = M(d, X(d))$
3. **ATE:** $\Delta = E[Y(1) - Y(0)]$ *note: the potential outcome can be rewritten as a function of both the treatment and the intermediate variable M.
4. **Total indirect effect**:

   $\delta^t(d) = E[Y(d, M(d, X(d)), X(d)) - Y(d, M(1-d, X(1-d)), X(d))],\ d \in \{0, 1\}$
5. **Partial indirect effect:**

   $\delta^p(d) = E[Y(d, M(d, X(d)), X(d)) - Y(d, M(1-d, X(d)), X(d))],\ d \in \{0, 1\}$
6. **Direct effect:**

   $\theta(d) = E[Y(d, M(d, X(d)), X(d)) - Y(1-d, M(d, X(d)), X(d))],\ d \in \{0, 1\},$

The direct effect is the change in the mean potential outcome due to an exogenous change in the treatment, while keeping the mediator and the covariates fixed. The direct effect here excludes channels via X. With that, $\theta(d)$, $\delta^p(d)$, and $\delta^t(d)$ do not add up to the ATE. Additionally, because Y (1,m) and Y(0,m) are treated as distinct quantities, the causal effect of the mediator M on Y can differ depending on the treatment status D. This directly allows for interaction between D and M in the model.

2.2 Assumptions

For the identifiability of direct and indirect effects in mediator analysis for our framework, a few crucial assumptions are necessary.

1. First is the random treatment assignment and conditional independence. Mathematically, this is written as:
   $\{Y(d, m, x), M(d, x)\} \perp D|X = x$ for all $d \in \{0, 1\}$ and m, x in support of M, X
   This assumption states that unconfoundedness of the treatment effects on the mediator and outcome must hold when conditioned on X. It implies that there are no unobserved confounders which affect both X and M and/or Y.
2. Second is conditional independence of the potential mediator state. Formally, this can be written as:
   $Y(d, m, X(d)) \perp M|D = d, X = x$ for all $d \in \{0, 1\}$ and m, x in the support of M, X, For obtaining comparable groups for any combination of covariate and mediator values, Pr $(D = d|M = m, X = x) > 0$ for all $d \in \{0, 1\}$ and m, x in the support of M, X is needed jointly.
3. Third is a functional form restriction w.r.t. potential mediators. This assumption states that the relationship between the Y and M given $D = d$ and $X(d)$ is the same functional form, regardless whether of the treatment state of M. This is key for the identification of $E[Y(d, M(1 - d, X(d)), X(d))]$ and thus the total indirect effect.
   $\mu_{d,x_d}(m_d) = E[Y(d, M(d, X(d)), X(d))|M(d, X(d)) = m_{1-d}, X(d) = x_d]$ for any m_d, x_d in the support of $M(d),\ X(d)$ and
   $\mu_{1-d,x_d}(m_{1-d}) = E[Y(d, M(1 - d, X(1 - d)), X(d))|M(1 - d, X(1 - d)) = m_{1-d}, X(d) = x_d]$
   for any m_{1-d}, x_d in the support of $M(1 - d),\ X(d)$.

2.3 Implications of Assumptions on Identifiability

Under assumptions 1 and 2, $\theta(d)$ relies on the identifiability of $E[Y(1 - d, M(d, X(d)), X(d))]$.

The average direct effect is now identified by:

$$\theta(d) = E[\frac{Y \cdot I\{D=1-d\}}{Pr(D=1-d|M,X)}] \cdot \frac{Pr(D=d|M,X)}{Pr(D=d)}] = E[(\frac{Y \cdot D}{Pr(D=1|M,X)} - \frac{Y \cdot (1-D)}{1-Pr(D=1|M,X)}) \cdot \frac{Pr(D=d|M,X)}{Pr(D=d)}]$$

Next, the partial indirect effect relies on the identification of $E[Y(d, M(1 - d, X(d)), X(d))]$:

$$\delta^p(d) = E[\frac{Y \cdot I\{D=d\}}{Pr(D=d)} - \frac{Y \cdot I\{D=d\}}{Pr(D=d|M,X)} \cdot \frac{Pr(D=1-d|M,X)}{Pr(D=1-d|X)} \cdot \frac{Pr(D=d|X)}{Pr(D=d)}].$$

Finally, under assumptions 1, 2, and 3, the average total indirect effect can be identified by:

$$\delta^t(d) = E[\frac{Y - \mu_{d,x}(E[M|D=1-d]) \cdot I\{D=d\}}{Pr(D=d)}]$$

Now, only the estimation of a (parametric or nonparametric) binary choice model for the propensity score can be plugged in.

3. Simulation Study and Empirical Application

3.1 Inverse Propensity Score Weighting

In our simulation, we applied IPW to estimate both the direct and indirect effects of treatment (D) on the outcome (Y) using data filtered specifically for females ($female = 1$). The outcome variable (Y) in our setting is $exhealth30$, which represents an individual's self-reported health status 30 months after assignment. The treatment (D) is $treat$, capturing exposure to an intervention, while the mediator (M) is $work2year2q$, indicating employment status one to one and a half years post-treatment. We account for pre-treatment covariates (X) such as education and training history, while a set of post-treatment confounders (W) includes employment status, vocational training participation, health indicators, and household characteristics.

After weighting, the sample balances pre-treatment covariates across treatment and control groups, allowing for the estimation of the direct effect by comparing the weighted outcomes between treated and untreated individuals. Similarly, the indirect effect is estimated by incorporating mediator propensity scores to account for selection into employment, ensuring that the observed mediator-outcome relationship is not confounded by post-treatment factors. Our approach refines Huber's method by carefully adjusting for mediator endogeneity and improving the stability of estimates through weight trimming, reducing sensitivity to extreme values and improving robustness in finite samples.

3.2 Alternative Estimation Methods

To test the robustness of the estimates of direct and indirect treatment effects using IPW, we employed three supplementary models, each addressing distinct challenges in causal mediation analysis. The Double Machine Learning (DML) approach leverages cross-fitting and doubly robust score functions to isolate mediation effects under selection-on-observables assumptions, combining post-LASSO models for treatment, mediator, and outcome estimation while trimming extreme probabilities to ensure numerical stability. Complementing this, the Doubly Robust (DR) estimator integrates inverse probability weighting with outcome regression, preserving consistency even if only the propensity score or outcome model is correctly specified, and decomposes effects into direct (e.g., E[Y(1,M(1))]−E[Y(0,M(1))] and indirect components (e.g., E[Y(1,M(1))]−E[Y(1,M(0))]), with bootstrapping for robust inference. Finally, Causal Forests (CF) extend this framework to heterogeneity analysis, using an ensemble of honest causal trees, trained via subsampling and EMSEτ-optimized splits, to partition the data into subgroups with distinct treatment effects, then decomposing total effects into direct and indirect pathways through a three-model sequence (total effect, mediator-adjusted, and mediator-as-treatment models). DML mitigates model misspecification through machine learning and orthogonality, DR enhances efficiency by combining weighting and regression, and CF uncovers nuanced subgroup-level mechanisms, ensuring comprehensive insights into causal pathways while addressing confounding, dimensionality, and effect heterogeneity.

3.3 Results

For both females and males, the total effects estimated by IPW are relatively small but positive. For females, the total effect is estimated at 0.0285 with a standard error of 0.0148, and the MSE is low (0.00108), indicating a relatively accurate estimate. The indirect effect for females is very small (0.00195), and its CI includes zero, suggesting it is not statistically significant. Similarly, for males, the total effect is 0.0219, with a slightly lower MSE (0.00037). However, the indirect effect for males appears more substantial (0.0173), with a 95% CI of [0.0055, 0.0296], indicating statistical significance and suggesting a meaningful mediation pathway. This suggests that, for males, a larger portion of the treatment effect is mediated through indirect channels compared to females.

The CF estimates show much higher MSEs, especially for the total and direct effects, on the order of 0.3 to 0.4, which suggests poor precision or higher variability across simulations. Nevertheless, the confidence intervals are very narrow and do not include zero, indicating strong statistical significance. For females, both direct and indirect effects are positive, with the indirect effect being 0.00465 (SE = 0.0153). For males, the indirect effect is again larger (0.0140) compared to females.

DML yields small effect sizes with very low MSEs for females, especially for the indirect effect (5.04e-07). For males, the MSE for indirect effect is slightly higher but still low. For females, the indirect effect is statistically significant with a narrow CI, suggesting that even small mediation effects are detectable with DML. For males, however, the indirect effect is not significant and close to zero. The precision of DML is high, especially in detecting subtle indirect effects for females.

The normal mediation approach provides moderate total and direct effects for both genders (ATE of 0.0258 for females and 0.00299 for males), with very small and statistically insignificant indirect effects, as their CIs include zero. This method may be less sensitive to detecting mediation compared to the other three.

A consistent pattern across methods is that males tend to have larger indirect effects than females, particularly evident in the IPW and CF results. This suggests that for males, a greater proportion of the total effect operates through indirect pathways. For females, the indirect effects are all close to 0, which suggests that employment does not seem to be a good mediator in measuring the effectiveness of the program. In contrast, direct effects for males are very small compared to females across all the models, which implies heterogeneous treatment effects across gender, which is well aligned with Huber's findings. Also, the direct effects for females are almost identical to the ATE across all the models, which suggests that the Job Corps program likely has a sizable direct effect that is not mediated by employment, which is also consistent with Huber's results.

4. Discussion

4.1 Discussion of Huber (2014)

The paper provides a flexible framework in causal mediation analysis and estimations without relying on strong parametric assumptions, making it especially valuable in observational studies where relationships are often complex and non-linear. By reducing dependence on specific functional form assumptions, Huber's approach generalises beyond methods like those of Flore and Flores-Lagunes (2009), which impose restrictive assumptions on potential outcomes, such as assuming the expected potential outcome as a linear function of the mediator. The method addresses selection bias and confounding in observational studies by leveraging on propensity score weighting, enhancing the credibility of causal interpretations. Furthermore, Huber advances the literature by discussing identification in more realistic scenarios where mediator exogeneity holds only conditional on post-treatment covariates, acknowledging that pretreatment variables may not fully capture endogeneity. This nuanced approach reflects real-world complexities, where treatments often influence variables that confound the mediator-outcome relationship.

Endogeneity in causal mediation analysis occurs when the mediator is correlated with unobserved confounded, leading to biased estimates of direct and indirect effects. The IPW estimator balances the distribution of observed confounders across treatment groups, but when mediators are endogenous, a naive application of IPW may not remove all biases due to omitted variables influencing both the mediator and outcome. Huber's control function approach models the residuals from the mediator equation and includes them in the weighting procedure. This corrects for potential endogeneity by controlling for unobserved confounders that affect both the mediator and outcome and partially corrects for misspecification in the mediator model (Wooldridge, 2007). By including the estimated residuals from the control function in the weighing model, Huber's method accounts for endogeneity, allowing for unbiased estimation of direct and indirect effects. His approach relaxes the strong assumption of sequential ignorability (Imai, Keele & Yamamoto, 2010) requiring no unobserved confounders affecting both the mediator and the outcome, which is often unrealistic (VanderWeele, T.J., 2015). By allowing for post-treatment confounders that potentially affects mediator-outcome relationships, Huber's mediation analysis is more robust to unmeasured confounding commonly found in observational studies.

While Huber's approach improves on traditional methods by addressing mediator endogeneity through a control function, it does not fully eliminate reliance on assumptions about unobserved confounders. The method relaxes sequential ignorability by including residuals from the mediator equation in the weighting procedure, but its validity hinges on the control function capturing all relevant confounding. However, Huber's approach does allow for some unobserved confounding, provided that it can be captured through the control function. If there are unobserved confounders that directly affect the outcome independently of the mediator, the control function cannot account for them. As such, domain expertise is still crucial to justify the plausibility of these assumptions.

In the event that the treatment of mediator assignment probabilities are misspecified due to incorrect functional form assumptions or omitted confounders, the resulting weights may fail to properly balance covariates, leading to biased estimates of direct and indirect effects (Khan & Tamer, 2010). Additionally, this method relies heavily on the common support assumption (Assumption 2(b)), requiring sufficient overlap in the distribution of covariates between treatment and control groups (Imbens, 2015). He quotes that if this assumption is violated, estimation becomes unstable and the variance of the estimates may explode. This is problematic especially in applications where the treatment or mediator assignment is highly selective, with propensity scores close to 0 or 1 (Lee, Lessler & Stuart, 2011). This leads to limited overlap between groups, extremely large weights for certain observations and high variability in the estimation process (Crump, Richard; Imbens, Guido, 2009). The IPW estimator potentially becomes highly sensitive to small changes in the data or propensity score model (Austin, Peter, 2011), making it difficult to draw reliable estimates and conclusions from the analysis. This issue is exacerbated in small samples, where the variability of estimated weights inflates the variance of treatment effect estimates, reducing statistical precision.

The paper lacks clear guidance and a data-driven method on aspects such as the selection of key hyperparameters, and bandwidth choices in kernel-based estimators or thresholds for trimming extreme propensity scores, which are crucial in ensuring the stability and reliability of estimated effects (Lee, Lessler & Stuart, 2011). Another constraint is its primary focus on binary treatments, making it less directly applicable to settings with multi-valued or continuous treatments without significant modifications (Imbens, 2000).

While IPW is a widely used method for estimating direct and total effects, Huber (2014) acknowledges that its application to indirect effects in mediation analysis remains underdeveloped and methodologically challenging. This gap is particularly problematic for policy and intervention studies, which often require a deeper understanding of mechanisms and causal pathways through which treatments exert their influence. Without a reliable framework for estimating indirect effects, researchers may struggle to disentangle direct treatment effects from mediation-driven effects, leading to incomplete causal conclusions. Adding on to the sensitive nature of IPW to accurately estimated propensity scores, the model must also account for the treatment-mediator-outcome relationship when estimating indirect effects, introducing an additional layer of complexity and potential misspecification (VanderWeele, 2015). This instability can make it difficult to derive precise and interpretable mediation effects, limiting the practical utility of IPW in mediation studies.

4.2 Discussion of Simulation Results

Given these challenges, our results extend Huber's framework by offering greater flexibility and robustness while addressing critical limitations in mediation analysis, particularly in estimating indirect effects. Huber's method primarily focuses on direct effects and does not provide a fully reliable framework for estimating indirect effects due to the compounding errors in the treatment-mediator and mediator-outcome relationships. This approach relies heavily on correctly specified propensity scores and struggles with mediator endogeneity. Our application of DML and DR estimators mitigates these issues through flexible modeling and cross-fitting. For instance, DML performed significantly better than IPW, demonstrating superior precision in detecting subtle indirect effects. Additionally, our decomposition of indirect effects via CF revealed gender-specific mediation pathways, with larger indirect effects for males, which Huber's binary treatment focus could not capture. By integrating machine learning and robustness checks, our study not only validates Huber's theoretical contributions but also provides a more reliable and nuanced toolkit for mediation analysis in complex, real-world settings.

Our application of DML for the estimation of indirect effects overcomes the parametric limitations inherent in Huber's approach. DML utilises machine learning algorithms, in our case, post-LASSO, to flexibly model both the treatment and outcome processes, mitigating the bias arising from incorrect functional form assumptions. Notably, DML outperformed all other models for both male and female subgroups. For female participants, DML achieves an MSE of just 0.00026 for total effects, nearly an order of magnitude lower than IPW's 0.00108 and dramatically better than CF's 0.317. This performance advantage holds for both direct and indirect effects, with DML's indirect effect estimates showing particular stability. The results highlight the flexibility of DML in capturing complex interactions that

IPW and traditional regression-based approaches fail to model accurately. Cross-fitting present also further enhances the robustness of DML, ensuring that overfitting does not inflate the variance of estimates, a critical consideration in high-dimensional or small-sample applications. The reduction in standard errors observed in DML estimates also implies greater statistical efficiency relative to IPW, suggesting that DML provides not only more reliable point estimates but also tighter inference intervals for direct and indirect effects. This is especially evident in the male subgroup results, where IPW produces large indirect effect estimates (0.0173) with concerning variability (SE = 0.0059), while DML yields stable, near-null estimates (-0.00115) that better align with theoretical expectations. Unlike Huber's control function approach, DML benefits from its double robustness and cross-fitting properties by combining flexible machine learning for nuisance functions with orthogonalised estimation, DML mitigates bias from misspecified models or high-dimensional confounders (Chernozhukov et al., 2018). This discrepancy underscores how IPW's reliance on a single model, making it vulnerable to finite-sample inefficiencies, while DML's hybrid approach of regression and weighting offers more reliable inference.

We also employ DR estimation to further improve robustness against misspecification. DR estimators combine IPW with outcome regression models, ensuring that if either the propensity score model or the outcome model is correctly specified, consistent estimates can still be obtained (Athey et al., 2024) . This is particularly relevant given the challenges identified in Huber's work, where estimation becomes unstable if the propensity score model is incorrectly specified. By incorporating DR methods for both direct and indirect effects, our approach mitigates the risk of relying solely on propensity scores, thereby improving precision and reducing the bias associated with poor overlap in covariates. Unlike Huber's approach, which assumes correct specification of the weighting model to achieve balance, DR methods allow for additional flexibility by modeling both the treatment-mediator and mediator-outcome relationships, ensuring greater robustness in mediation analysis. For female subjects, DR achieves a balanced direct effect estimate (0.0282, SE = 0.0159) with tighter confidence bounds compared to CF (0.0264, SE = 1.845), highlighting its resilience to extreme variance. In males, DR's direct effect (0.0005, SE = 0.0150) shows greater stability than IPW (0.00227, SE = 0.0187) and avoids the volatile variance seen in CF (0.0196, SE = 9.348), likely stemming from CF's sensitivity to small-sample noise or model misspecification in heterogeneous subgroups. This underscores DR's ability to deliver more consistent and efficient direct effect estimates by leveraging both weighting and regression adjustments.

As for Causal Forest, the results highlight several points that contrast with the performance of DML, DR and IPW. While, CF was able to detect heterogeneous treatment effects, being able to identify subgroup variance (female ATE = 0.0262 , male ATE = 0.0182), it's performance also consistently produces significantly higher MSEs, with the female subset having an MSE of 0.3179 and the male subset having an MSE of 0.4094. Notably, CF's indirect effects under treatment differ substantially between genders (0.00465 for females, 0.0140 for males), highlighting its capacity to detect heterogeneity but also its susceptibility to sample-specific noise. While CF's are powerful for capturing treatment effect heterogeneity and estimating Conditional Average Treatment Effects (CATEs), their performance was suboptimal with unusually high standard errors and MSEs. This is likely because the objective of Huber's framework, and our evaluation, was to estimate ATEs, not individual-level variation. CF, by design, emphasises local effects and partition the data into subgroups to estimate heterogeneous responses. As a result, they can introduce instability and noise when aggregated for ATE inference, especially in small samples or when mediation effects are subtle. This mismatch in estimation goals likely explains the poor performance and large variance we observed. Additionally, CF appears prone to overfitting, likely due to its complex ensemble approach. Overfitting in CF might be particularly problematic in cases where the depth or number of trees are not properly tuned, leading to the high variance in the estimated treatment effects and large MSEs. These limitations highlight that the model's flexibility requires careful tuning of the tree depth and sample-splitting parameters.

4.3 Limitations of Simulation

Across all models, the consistently greater stability in the estimates of female participants as compared to the poor performance of models in the male subgroup might indicate the presence of potential unmodelled effect modifiers or underlying heterogeneity across genders. These differences may arise from structural disparities in the data generating process, such as varying levels of noise, different treatment effects, or unequal amounts of confounding, that our simulation might have failed to capture. We also relied heavily on the observed covariates to adjust for post-treatment confounding, which may be insufficient if time-varying unobserved confounders are excluded or incorrectly specified. By assuming homogeneity and failing to explicitly model these modifiers, we risk producing biased or unstable estimates in subgroups where these assumptions do not hold, particularly in this context for males.

Additionally, while we incorporate supplementary models (DML, DR, CF) to address various sources of bias and model misspecification, each method has its own vulnerabilities. Causal Forest for instance, despite uncovering treatment effect heterogeneity, exhibits unusually high MSEs and narrow confidence intervals, which alludes to possible overfitting or an overly complex model structure relative to the sample size. Across methods, our reliance on estimated propensity scores and trimming procedures also introduces bias-variance trade-offs that may obscure true effects.

4.4 Reflection and Future Areas of Exploration

From a methodological standpoint, while this study evaluates the findings of Huber against several supplementary models, further works should focus explicitly on modelling effect heterogeneity and identifying potential effect modifiers. Stratified or interaction-based models, built upon the frameworks used in our study could help uncover structural disparities (Song & Kawai, 2022) in the data generating process that our methods fail to capture. Additionally, incorporating methods that better address unobserved and time-varying confounders (Zhang, Zhang, & Zhou, 2023), such as instrumental variables, could enhance the validity of post-treatment adjustments made in IPW and DML

Through this project, our group found that the computational intensity to run and train the models was a big challenge. It made tuning, robustness checks, and replication extremely time consuming. Since the bulk of our discussion relied heavily on the model outputs, the long run times significantly limited the time available for deeper exploration and interpretation of the results, limiting our ability to fully investigate patterns or anomalies that may have further enriched the analysis. To overcome this challenge, our group decreased the number of bootstrap samples and split the models within ourselves such that each person would run 1-2 models. By tuning the parameters and delegating the workload of running the models, we saved a substantial amount of time and were able to move on with the rest of the project. Furthermore, the inability for some models, like Causal Forest, to produce results under trimming limits our ability to compare the outputs across models, especially when dealing with outlier-prone samples like this one. Finally, as mediation analysis was a relatively new concept for us, there was an initial learning curve in adapting the causal inference tools we had learned throughout the course to estimate indirect effects within a mediation framework.

5. Conclusion

This study explored the application of Huber’s IPW and supplementary methods to estimate direct and indirect effects in causal mediation analysis. The findings demonstrated that while IPW provides a flexible framework for mediation analysis, its reliance on propensity score accuracy and common support assumptions can lead to instability, particularly in estimating indirect effects. DML emerged as a robust alternative, offering lower MSEs and greater precision, especially for detecting subtle mediation effects.

Overall, this study contributes to advancing mediation analysis by integrating modern machine learning methods with traditional causal frameworks, offering actionable insights for policy and intervention studies.

6. References

Austin, P. C. (2011). An introduction to propensity score methods for reducing the effects of confounding in observational studies. Multivariate Behavioral Research, 46(3), 399–424. https://doi.org/10.1080/00273171.2011.568786

Athey, S., Bayati, M., Doudchenko, N., Imbens, G., & Khosravi, K. (2024). Doubly robust inference in causal latent factor models. MIT Department of Economics. https://economics.mit.edu/sites/default/files/2024-02/Doubly%20Robust%20Inference%20in%20Causal%20Latent%20Factor%20Models.pdf

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. https://doi.org/10.1111/ectj.12097

Crump, R. K., Hotz, V. J., Imbens, G. W., & Mitnik, O. A. (2009). Dealing with limited overlap in estimation of average treatment effects. Biometrika, 96(1), 187–199. https://doi.org/10.1093/biomet/asn055

Flores, C. A., & Flores-Lagunes, A. (2009). Identification and estimation of causal mechanisms and net effects of a treatment under unconfoundedness. https://docs.iza.org/dp4237.pdf

Huber, M. (2014). Identifying causal mechanisms (primarily) based on inverse probability weighting. Journal of Applied Econometrics,29(6), 920-943. https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2341https://onlinelibrary.wiley.com/doi/abs/10.1002/jae.2341

Imai, K., Keele, L., & Yamamoto, T. (2010). Identification, inference, and sensitivity analysis for causal mediation effects. Statistical Science, 25(1), 51–71. https://doi.org/10.1214/10-STS321

Imbens, G. W. (2015). Matching methods in practice: Three examples. Journal of Human Resources, 50(2), 373–419. https://doi.org/10.3368/jhr.50.2.373

Imbens, G. W. (2000). The role of the propensity score in estimating dose-response functions. Biometrika, 87(3), 706–710. https://doi.org/10.1093/biomet/87.3.706

Kaplan, D. (2000). Structural equation modeling: Foundations and extensions. Sage Publications. https://www.researchgate.net/publication/233337570_Book_review_of_D_Kaplan_2000_Structural_equation_modeling_Foundations_and_extensions

Kang, J. D. Y., & Schafer, J. L. (2007). Demystifying double robustness: A comparison of alternative strategies for estimating a population mean from incomplete data. Statistical Science, 22(4), 523–539. https://doi.org/10.1214/07-STS227

Khan, S., & Tamer, E. (2010). Inference on endogenously censored regression models using conditional moment inequalities. Journal of Econometrics, 155(1), 1–13. https://doi.org/10.1016/j.jeconom.2009.09.006

Lee, B. K., Lessler, J., & Stuart, E. A. (2011). Weight trimming and propensity score weighting. PLOS ONE, 6(3), e18174. https://doi.org/10.1371/journal.pone.0018174

Song, C., & Kawai, R. (2022). Understanding the causal effect of interest rates on the housing market: Evidence from Japan. Journal of Macroeconomics, 72, 103418. https://doi.org/10.1016/j.jmacro.2022.103418

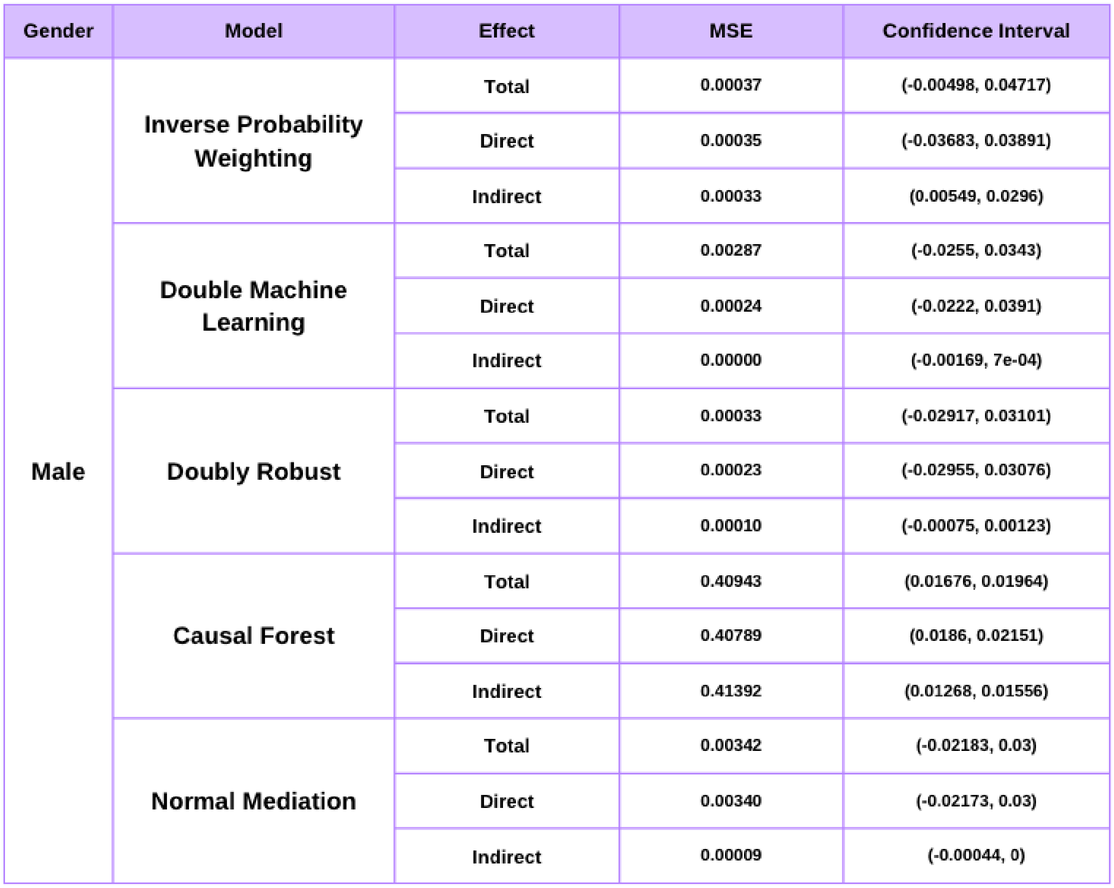
VanderWeele, T. J. (2015). Explanation in causal inference: Methods for mediation and interaction. Oxford University Press. https://psycnet.apa.org/record/2015-10736-000

Wooldridge, J. M. (2007). Econometric analysis of cross section and panel data (2nd ed.). MIT Press. https://mitpress.mit.edu/9780262232586/econometric-analysis-of-cross-section-and-panel-data/

Zhang, L., Zhang, J., & Zhou, X. (2023). A novel approach to causal inference with unobserved confounding. arXiv. https://arxiv.org/abs/2312.07175

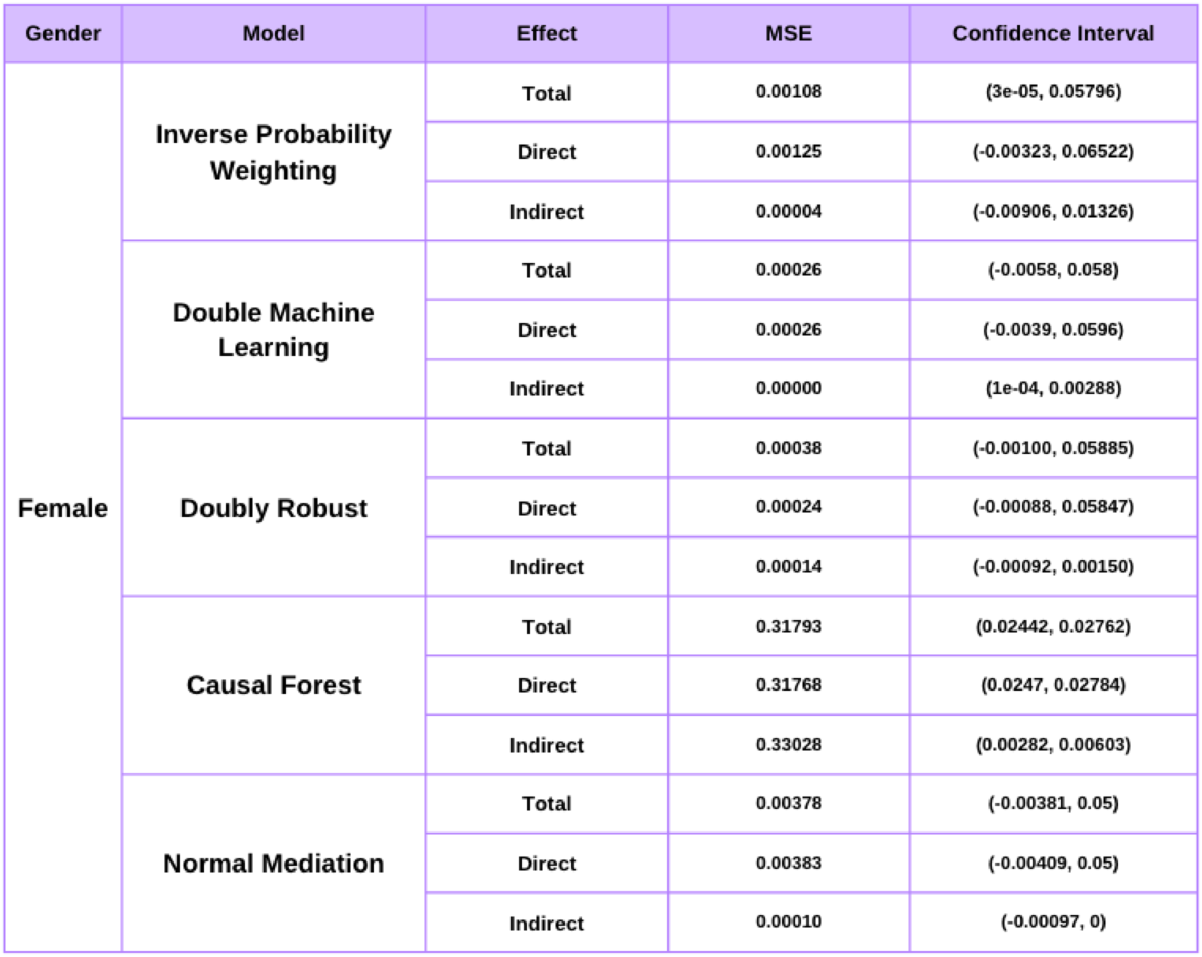
Appendix

The replication code files and dataset can be found at https://github.com/yangshutingg/dse4231



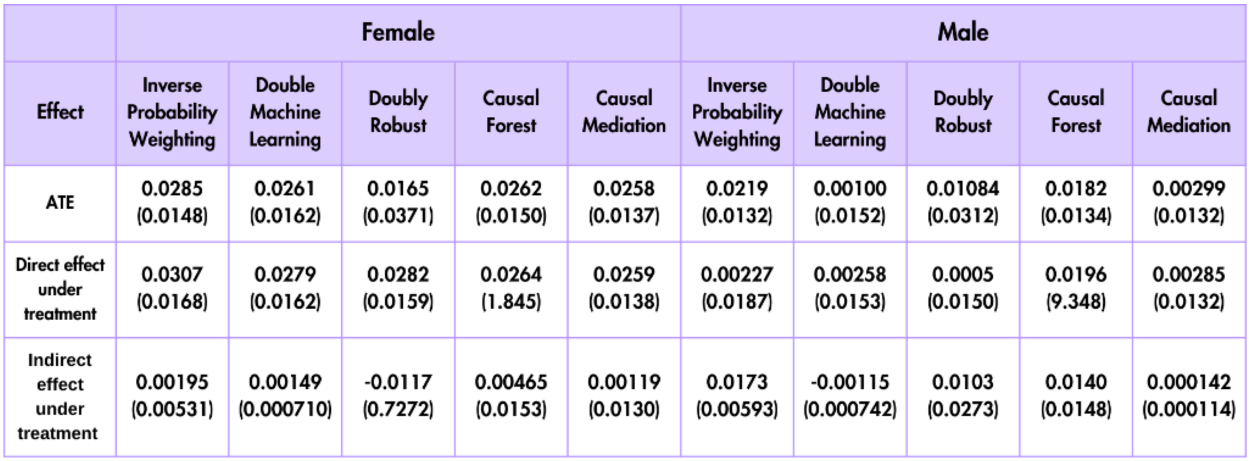
Gender	Model	Effect	MSE	Confidence Interval
Male	Inverse Probability Weighting	Total	0.00037	(-0.00498, 0.04717)
		Direct	0.00035	(-0.03683, 0.03891)
		Indirect	0.00033	(0.00549, 0.0296)
	Double Machine Learning	Total	0.00287	(-0.0255, 0.0343)
		Direct	0.00024	(-0.0222, 0.0391)
		Indirect	0.00000	(-0.00169, 7e-04)
	Doubly Robust	Total	0.00033	(-0.02917, 0.03101)
		Direct	0.00023	(-0.02955, 0.03076)
		Indirect	0.00010	(-0.00075, 0.00123)
	Causal Forest	Total	0.40943	(0.01676, 0.01964)
		Direct	0.40789	(0.0186, 0.02151)
		Indirect	0.41392	(0.01268, 0.01556)
	Normal Mediation	Total	0.00342	(-0.02183, 0.03)
		Direct	0.00340	(-0.02173, 0.03)
		Indirect	0.00009	(-0.00044, 0)

Table 1: MSE and CI Widths for Male Subgroup



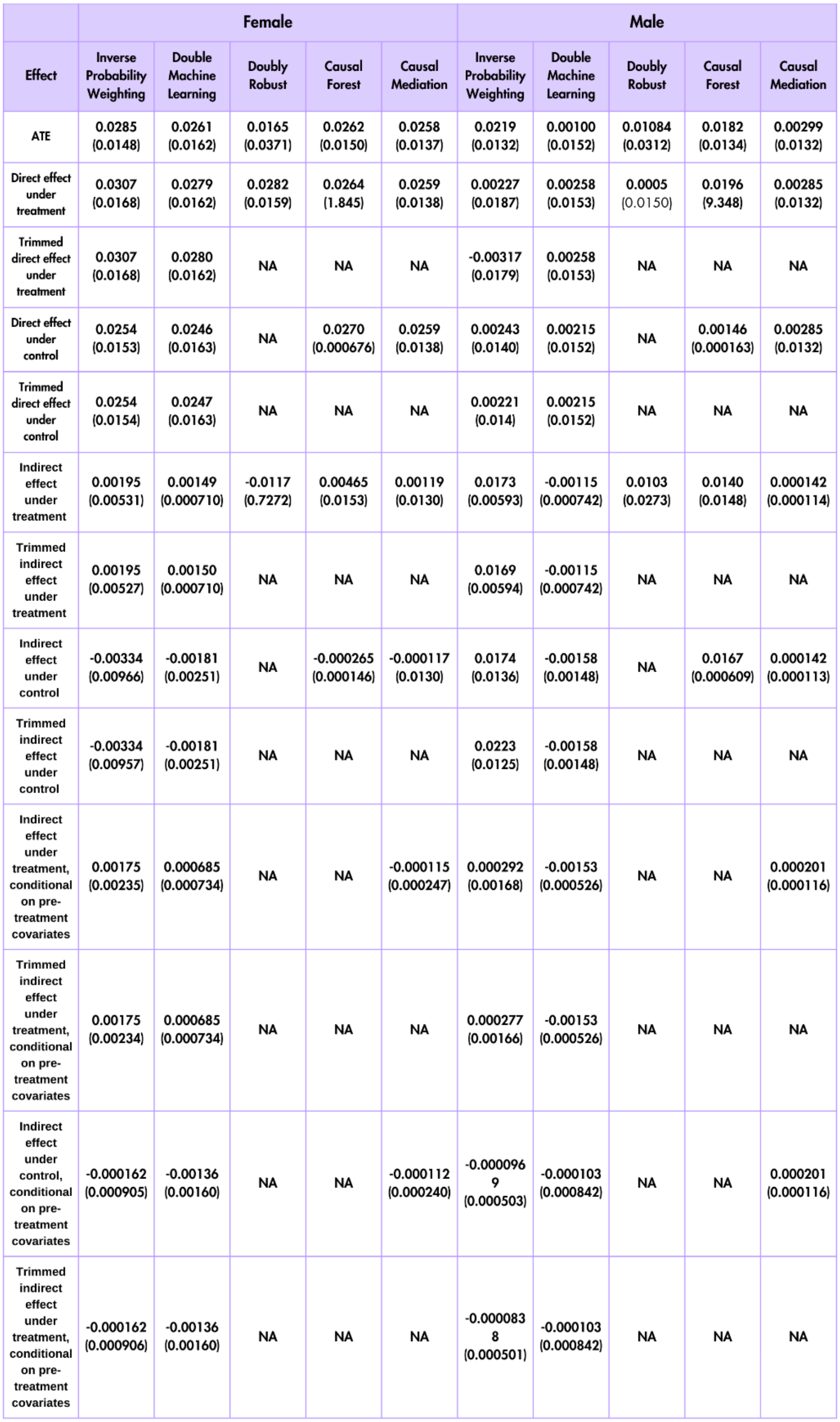
Gender	Model	Effect	MSE	Confidence Interval
Female	Inverse Probability Weighting	Total	0.00108	(3e-05, 0.05796)
		Direct	0.00125	(-0.00323, 0.06522)
		Indirect	0.00004	(-0.00906, 0.01326)
	Double Machine Learning	Total	0.00026	(-0.0058, 0.058)
		Direct	0.00026	(-0.0039, 0.0596)
		Indirect	0.00000	(1e-04, 0.00288)
	Doubly Robust	Total	0.00038	(-0.00100, 0.05885)
		Direct	0.00024	(-0.00088, 0.05847)
		Indirect	0.00014	(-0.00092, 0.00150)
	Causal Forest	Total	0.31793	(0.02442, 0.02762)
		Direct	0.31768	(0.0247, 0.02784)
		Indirect	0.33028	(0.00282, 0.00603)
	Normal Mediation	Total	0.00378	(-0.00381, 0.05)
		Direct	0.00383	(-0.00409, 0.05)
		Indirect	0.00010	(-0.00097, 0)

Table 2: MSE and CI Widths for Female Subgroup



	Female					Male				
Effect	Inverse Probability Weighting	Double Machine Learning	Doubly Robust	Causal Forest	Causal Mediation	Inverse Probability Weighting	Double Machine Learning	Doubly Robust	Causal Forest	Causal Mediation
ATE	0.0285 (0.0148)	0.0261 (0.0162)	0.0165 (0.0371)	0.0262 (0.0150)	0.0258 (0.0137)	0.0219 (0.0132)	0.00100 (0.0152)	0.01084 (0.0312)	0.0182 (0.0134)	0.00299 (0.0132)
Direct effect under treatment	0.0307 (0.0168)	0.0279 (0.0162)	0.0282 (0.0159)	0.0264 (1.845)	0.0259 (0.0138)	0.00227 (0.0187)	0.00258 (0.0153)	0.0005 (0.0150)	0.0196 (9.348)	0.00285 (0.0132)
Indirect effect under treatment	0.00195 (0.00531)	0.00149 (0.000710)	-0.0117 (0.7272)	0.00465 (0.0153)	0.00119 (0.0130)	0.0173 (0.00593)	-0.00115 (0.000742)	0.0103 (0.0273)	0.0140 (0.0148)	0.000142 (0.000114)

Table 3: Main Estimates and their standard errors for Male and Female Subgroups



	Female					Male				
Effect	Inverse Probability Weighting	Double Machine Learning	Doubly Robust	Causal Forest	Causal Mediation	Inverse Probability Weighting	Double Machine Learning	Doubly Robust	Causal Forest	Causal Mediation
ATE	0.0285 (0.0148)	0.0261 (0.0162)	0.0165 (0.0371)	0.0262 (0.0150)	0.0258 (0.0137)	0.0219 (0.0132)	0.00100 (0.0152)	0.01084 (0.0312)	0.0182 (0.0134)	0.00299 (0.0132)
Direct effect under treatment	0.0307 (0.0168)	0.0279 (0.0162)	0.0282 (0.0159)	0.0264 (1.845)	0.0259 (0.0138)	0.00227 (0.0187)	0.00258 (0.0153)	0.0005 (0.0150)	0.0196 (9.348)	0.00285 (0.0132)
Trimmed direct effect under treatment	0.0307 (0.0168)	0.0280 (0.0162)	NA	NA	NA	-0.00317 (0.0179)	0.00258 (0.0153)	NA	NA	NA
Direct effect under control	0.0254 (0.0153)	0.0246 (0.0163)	NA	0.0270 (0.000676)	0.0259 (0.0138)	0.00243 (0.0140)	0.00215 (0.0152)	NA	0.00146 (0.000163)	0.00285 (0.0132)
Trimmed direct effect under control	0.0254 (0.0154)	0.0247 (0.0163)	NA	NA	NA	0.00221 (0.014)	0.00215 (0.0152)	NA	NA	NA
Indirect effect under treatment	0.00195 (0.00531)	0.00149 (0.000710)	-0.0117 (0.7272)	0.00465 (0.0153)	0.00119 (0.0130)	0.0173 (0.00593)	-0.00115 (0.000742)	0.0103 (0.0273)	0.0140 (0.0148)	0.000142 (0.000114)
Trimmed indirect effect under treatment	0.00195 (0.00527)	0.00150 (0.000710)	NA	NA	NA	0.0169 (0.00594)	-0.00115 (0.000742)	NA	NA	NA
Indirect effect under control	-0.00334 (0.00966)	-0.00181 (0.00251)	NA	-0.000265 (0.000146)	-0.000117 (0.0130)	0.0174 (0.0136)	-0.00158 (0.00148)	NA	0.0167 (0.000609)	0.000142 (0.000113)
Trimmed indirect effect under control	-0.00334 (0.00957)	-0.00181 (0.00251)	NA	NA	NA	0.0223 (0.0125)	-0.00158 (0.00148)	NA	NA	NA
Indirect effect under treatment, conditional on pre-treatment covariates	0.00175 (0.00235)	0.000685 (0.000734)	NA	NA	-0.000115 (0.000247)	0.000292 (0.00168)	-0.00153 (0.000526)	NA	NA	0.000201 (0.000116)
Trimmed indirect effect under treatment, conditional on pre-treatment covariates	0.00175 (0.00234)	0.000685 (0.000734)	NA	NA	NA	0.000277 (0.00166)	-0.00153 (0.000526)	NA	NA	NA
Indirect effect under control, conditional on pre-treatment covariates	-0.000162 (0.000905)	-0.00136 (0.00160)	NA	NA	-0.000112 (0.000240)	-0.0000969 (0.000503)	-0.000103 (0.000842)	NA	NA	0.000201 (0.000116)
Trimmed indirect effect under treatment, conditional on pre-treatment covariates	-0.000162 (0.000906)	-0.00136 (0.00160)	NA	NA	NA	-0.0000838 (0.000501)	-0.000103 (0.000842)	NA	NA	NA

Table 4: Estimates with Trimming for Female and Male Subgroups

Double Machine Learning (DML)

We adopt the mediation DML (medDML) approach using the causalweight package, which estimates natural direct and indirect effects under a selection-on-observables assumption. Specifically, we assume that all confounders of the treatment, mediator, and outcome relationships are observed and unaffected by the treatment. This doubly robust estimator leverages recent advances in double machine learning (Chernozhukov et al., 2018) and efficient score functions (Tchetgen Tchetgen and Shpitser, 2012; Farbmacher et al., 2019) to provide robust estimates of causal mediation effects.

We first specify pre-treatment covariates X, which are assumed to jointly affect the treatment D, mediator M, and outcome Y. The method begins by estimating the treatment and mediator models using post-LASSO logistic regression, and the outcome model using post-LASSO regression, implemented using rlassologit and rlasso functions from the hdm package. These models are trained using cross-fitting with default 3-fold splitting, whereby the sample is divided into folds and model parameters are learned in one fold and used to predict in another. This guards against overfitting and ensures orthogonality between nuisance parameter estimation and causal effect estimation.

The procedure estimates the conditional expectations and probabilities required for efficient score computation. For stability, we implement a trimming rule, excluding observations with extremely small estimated probabilities (by default, $trim = 0.5$) to avoid division by near-zero terms in the estimation of potential outcomes.

In the second stage, we use the learned nuisance models to predict the efficient scores for natural direct and indirect effects. By averaging these scores across the sample (with roles of folds swapped), we obtain consistent estimates of the causal mechanisms. This estimation is doubly robust, meaning that it remains valid if either the model for the outcome or the models for the treatment and mediator are correctly specified.

Finally, standard errors are derived using asymptotic approximations, based on the variance of the estimated efficient scores. This accounts for the uncertainty from both the causal effect estimation and the machine learning-based nuisance models.

This DML-based mediation framework provides several advantages over traditional methods such as IPW or standard regression-based mediation, allowing for high-dimensional control variables, mitigating model misspecification risks, and offering robustness through cross-fitting and doubly robust score construction.

Doubly Robust (DR)

DR estimators further enhance the robustness of causal effect estimation by combining IPW with outcome regression modeling. Unlike standard IPW, which relies solely on correctly estimating the propensity score, DR estimators remain consistent if either the propensity score model or the outcome model is correctly specified. This dual robustness property makes them particularly useful in mediation settings, where misspecification of either model can lead to bias. Applied to our dataset, the DR approach involves first estimating the propensity scores for treatment and mediator assignment, as in the IPW method. Simultaneously, we fit regression models for the outcome (Y) as a function of the treatment (D), mediator (M), and covariates (X, W). The final DR estimator combines these two components, adjusting the outcome regression using inverse probability weights to ensure proper balancing.

In our implementation, the mediation function first estimates the propensity scores using probit regression models. The treatment probability given the mediator, post-treatment confounders, and pre-treatment confounders is estimated using:

$$P(D = 1 \mid M, W, X)$$

Additionally, alternative propensity score estimates are generated based on the different subset of covariates, such as $P(D = 1|X)$, $P(D = 1|W, X)$ and $P(D = 1|M, X)$. These propensity scores are then used to compute the IPWs required for DR estimation.

Next, we estimate the outcome regression models separately for treated and control groups usingOLD. The predicted outcomes under different mediator and treatment assignments are then computed. The direct effect under treatment and control conditions is calculated as:

$DE_{treat} = E[Y(1,\ M(1))] - E[Y(0,\ M(1))]$ and $DE_{control} = E[Y(1,\ M(0))] - E[Y(0,\ M(0))]$

Similarly, the indirect effects capture how changes in the mediator affect the outcome under treatment and control conditions:

$IE_{treat} = E[Y(1,\ M(1))] - E[Y(1,\ M(0))]$ and $IE_{control} = E[Y(0,\ M(1))] - E[Y(0,\ M(0))]$

To improve robustness, we further decompose the indirect effects into total and partial effects based on different assumptions regarding mediator-treatment interactions:

$IE_{total} = E[Y(1,\ M(1))] - E[Y(0,\ M(0))]$ and $IE_{partial} = E[Y(1,\ M(1))] - E[Y(1,\ M\ ^*)]$

where M* represents a counterfactual mediator value adjusted for post-treatment confounding. We also then implement bootstrapping to estimate standard errors for the treatment, direct, and indirect effects.

By incorporating both weighting and regression adjustments, DR estimators provide an additional layer of robustness, mitigating the sensitivity of our estimates to extreme weights and functional form assumptions. Compared to Huber's standard IPW approach, DR offers more reliable effect estimation in cases where propensity scores are misspecified or suffer from limited common support. This is particularly relevant for our study, where the treatment and mediator assignment mechanisms are complex and may not be fully captured by a single parametric model.

Casual Forest (CF)

Causal forests are a machine learning method designed to estimate heterogeneous treatment effects, how the impact of a treatment varies across different subgroups, by extending decision tree and random forest frameworks. Originally, decision trees were used for classification and regression by partitioning data based on covariates to predict an outcome. However, these trees often suffered from high variance and overfitting, especially when grown deep to fit training data exactly, minimising its effectiveness in uncovering causal effects in general. Athey and Imbens (2016) addressed this gap with causal trees, adapting the tree framework to estimate the conditional average treatment effect (CATE) by introducing a new splitting criterion. This allows trees to split based on treatment effect heterogeneity rather than predictive accuracy. Casual trees utilise honest estimation, splitting data into two parts; one for determining the tree structure (training) and the other for estimating treatment effects (estimation). This ensures that estimates remain unbiased and valid for inference.

However, relying on a single causal tree retained the issue of high variance. Wager and Athey (2018) developed Causal Forests, an ensemble of causal trees built using subsampling without replacement. CFs preserve honest estimation and use small, deep tree leaves to capture finer heterogeneity. They also incorporate adaptive local weighting, such that the CATE estimate for a point is influenced more by similar observations, akin to nearest neighbours. This improves stability and ensures that treatment effect estimates are not driven by outliers or specific trees. CFs also offer asymptotic normality, enabling valid statistical inference.

The mechanism of CFs utilises this process:

1. Honest estimation: randomly split the dataset (if testing is done, then this dataset would be the training dataset) into two, one for training and one for estimation.
2. We start with the entire training dataset as a single root node, where both treated and untreated observations are included and iterate based on depth limit or stopping criteria

3. For each depth in the maximum depth, iterate over the covariates. For each covariate and each possible split point; (eg. age < 30 and age > 30)

3.1 - Divide the data into two branches based on the chosen covariates and split points

3.2 - Within each subgroup/branch, we compute the CATE

3.3 - Calculate the modified splitting criterion, Expected Mean Squared Error for Treatment effects (*EMSEτ*)

4. Choose the best split for that covariate by minimising the *EMSEτ*

5. When the tree stops growing, assign each observation in the estimation dataset to the corresponding leaf. Since each leaf is fairly large, to estimate the treatment effect by taking the difference in means between the treated and untreated groups; propensity score weighting is applied to correct

In our implementation, we employed a three-model CF framework, leveraging on the flexibility of its methodology. The first model estimates the total treatment effect by training a CF with estimates X = x, outcome Y = y, and treatment indicator W = d. The second model estimates the direct effect by controlling for the mediator. This involves a two step process: by first training a separate CF to model the relationship between the treatment and mediator (X = x, Y = m, W = d) and use the resulting predicted mediator values to include as an additional covariate in a new CF model, with X = x, Y = y and W = d. The third model estimates the indirect effect by training a CF where the mediator m replaces the original treatment variable (X = x, Y = y, W = m). The direct effect on the treated is then taken from the second model's predictors and the indirect effect on the treated is derived from the third model. We also estimate direct and indirect effects on the control group through:

$$DirectEffect_{control} = DirectEffect_{treated} - TotalEffect$$

$$IndirectEffect_{control} = TotalEffect - DirectEffect_{treated}$$

This decomposition of effects allows for a nuanced understanding of the pathways through which treatment influences outcomes and serves as a robust method to complement the IPW strategy.

Figure 1: Causal forest plot for females

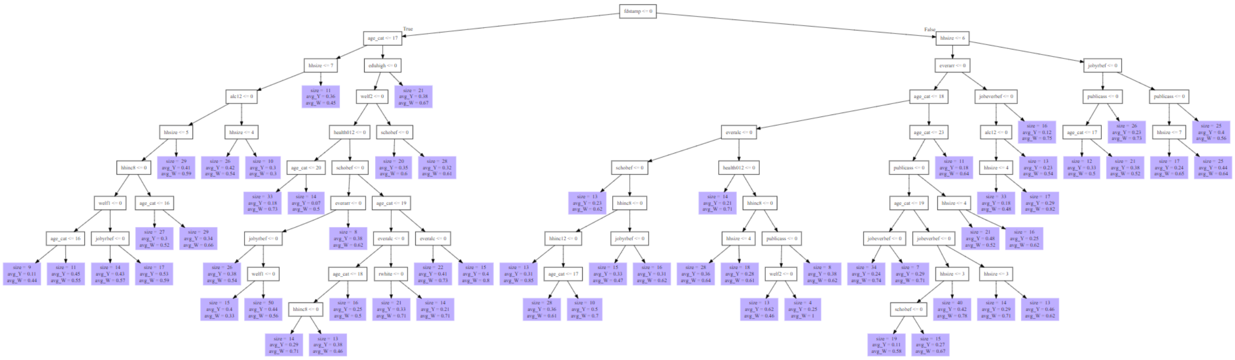


Figure 2: Causal forest plot for males

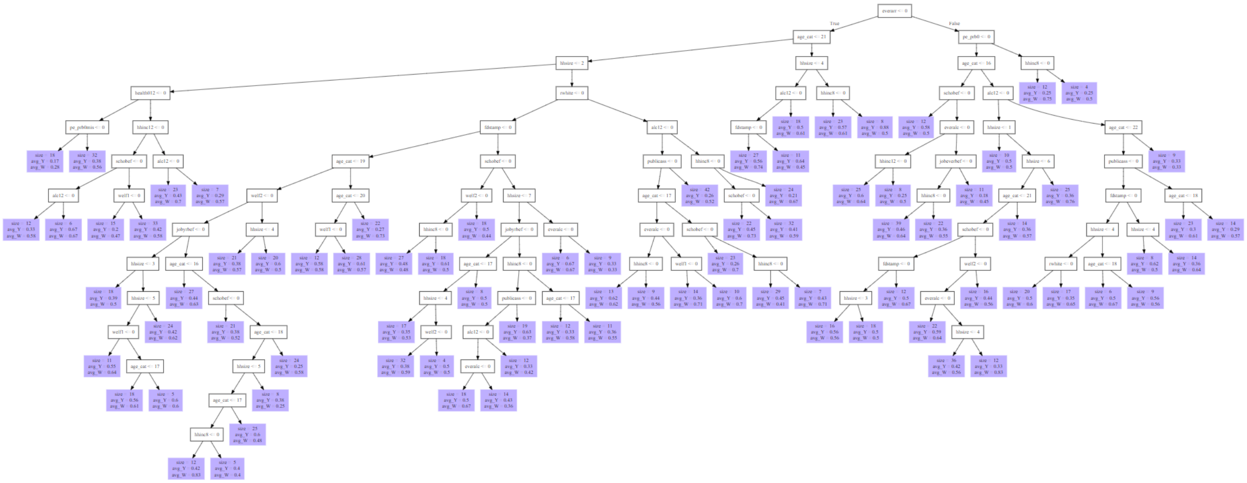


Figure 3: Propensity score under treatment for females

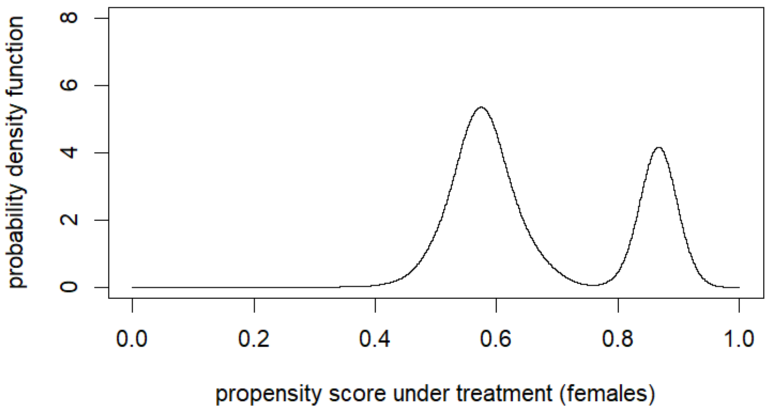


Figure 4: Propensity score under non-treatment for females

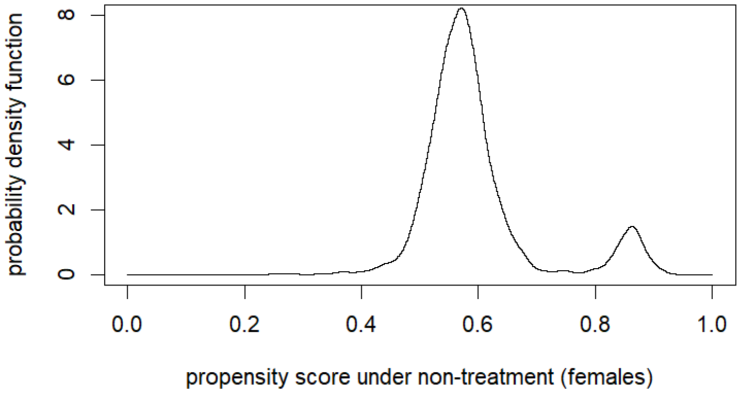


Figure 5: Propensity score under treatment for males

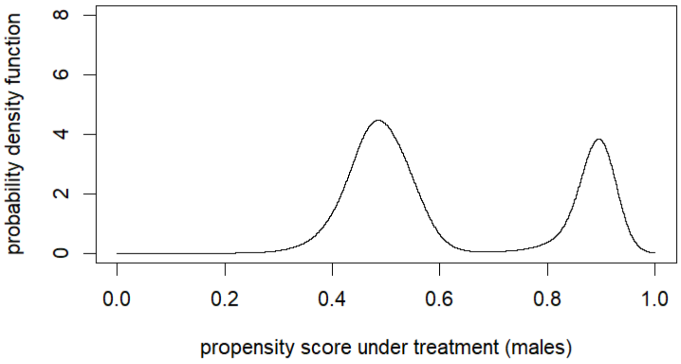
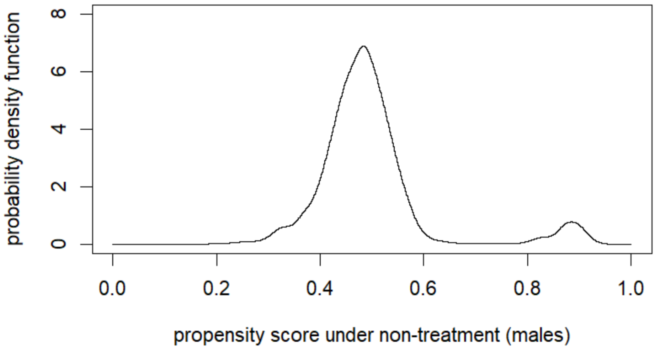


Figure 6: Propensity score under non-treatment for males


probability density function
propensity score under non-treatment (males)