

# Summary of ATE Estimators: Pros and Cons

From Whiteboard Session

This document summarizes the pros and cons of various Average Treatment Effect (ATE) estimators as transcribed from the whiteboard screenshots.

Estimator	Pros	Cons
<b>Linear Regression Adjustment</b>	<ul style="list-style-type: none"><li>• Simplicity</li><li>• No tuning required</li><li>• Interpretable</li><li>• Provides confidence intervals (CIs)</li><li>• Very precise if the Conditional Expectation Function (CEF) is truly linear</li></ul>	<ul style="list-style-type: none"><li>• Too simple; likely misspecified</li><li>• Assumes linear CEF</li><li>• Problematic if overlap is poor (extrapolation)</li></ul>
<b>G-formula (S-learner)</b>	<ul style="list-style-type: none"><li>• Can capture non-linearities</li></ul>	<ul style="list-style-type: none"><li>• Slow convergence rate</li><li>• Can be too confident (overfitting) if not cross-fit</li><li>• No straightforward confidence intervals</li></ul>
<b>G-formula (T-learner)</b>	<ul style="list-style-type: none"><li>• Often better bias properties than S-learner</li><li>• Does not “regularize” the treatment</li></ul>	<ul style="list-style-type: none"><li>• Can have worse accuracy for the minority treatment group in imbalanced datasets</li><li>• No straightforward confidence intervals</li></ul>

Estimator	Pros	Cons
IPW (ML)	<ul style="list-style-type: none"> <li>• Only requires learning the propensity score, not the outcome model</li> <li>• Good accuracy if treatment assignment mechanism is simple and easy to learn from data</li> </ul>	<ul style="list-style-type: none"> <li>• High variance</li> <li>• Very sensitive to poor overlap (extreme propensity scores)</li> <li>• No straightforward confidence intervals</li> </ul>
IPW (Logistic)	<ul style="list-style-type: none"> <li>• Simpler than ML-based IPW</li> <li>• Provides confidence intervals (CIs) (through delta method; which we have not seen in class)</li> </ul>	<ul style="list-style-type: none"> <li>• Assumes a simple (logistic) form for the propensity score model</li> </ul>
Doubly Robust (DR)	<ul style="list-style-type: none"> <li>• <b>Double robustness:</b> Consistent if either the outcome or propensity model is correct</li> <li>• Lower variance than IPW</li> <li>• Asymptotically normal even though ML is used for estimation</li> <li>• Provides confidence intervals (CIs) even though ML is used</li> <li>• Insensitive to errors of ML models</li> </ul>	<ul style="list-style-type: none"> <li>• Cannot empirically test the product rate condition for convergence</li> <li>• More randomness due to k-fold cross-validation (can be ameliorated by repeating the process multiple times with different k-folds and taking the median estimate and the median standard error across runs)</li> <li>• Still suffers from substantial (albeit unavoidable without further assumptions) variance if there are covariate regions with low overlap (extreme propensities)</li> </ul>

Estimator	Pros	Cons
<b>DR (Semi-crossfitting)</b>	<ul style="list-style-type: none"> <li>• Better computational performance</li> <li>• Better for tuning</li> </ul>	<ul style="list-style-type: none"> <li>• Less guardrails against overfitting</li> <li>• If you overdo it with tuning (i.e. let automl run for too long and examine too many models), then you are susceptible to some bias due to overfitting.</li> </ul>
<b>DR (Stacked-semi-crossfitting)</b>	<ul style="list-style-type: none"> <li>• Reduce variance of tuning</li> <li>• Improve RMSE performance of outcome and propensity model, due to ensembling</li> </ul>	<ul style="list-style-type: none"> <li>• Even less guardrails against overfitting</li> <li>• You have to be careful to stack (ensemble) only among a small set of models (e.g. best model of each model type, e.g. (best random forest, best gradient boosted forest, best penalized linear model))</li> </ul>