Propensity Score Estimation

WALTER LEITE

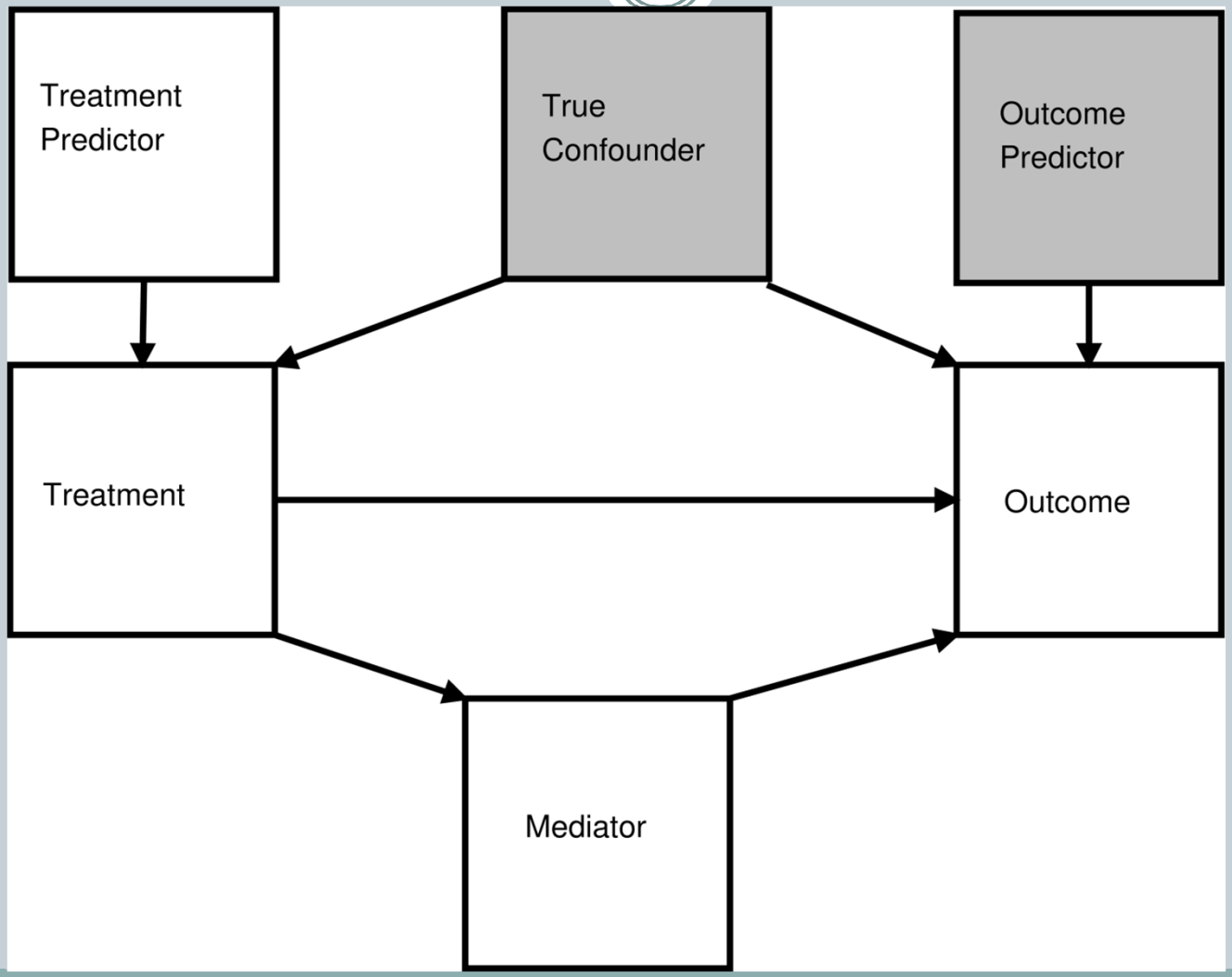
Steps

Step	Description	Example Methods
Selection of covariates	Identify true confounders and predictors of outcome	Include pretest score or outcome proxies
Dealing with missing data	Use available data or impute missing data	Listwise deletion, single imputation, multiple imputation
Propensity score estimation	Estimate predicted probability of treatment assignment	Logistic regression, Random forests, eneralized boosted modeling
Evaluation of common support	Compare distributions of propensity scores	Histograms, Box-and-whisker plots, Kernel density plots

Example – Effect of Career Academy on Income

- Research question: Does participating in a career academy in high school affect future income?
- Data source: Education Longitudinal Study (ELS)
- Sampling design: two-stage stratified sampling method where schools were sampled with probability proportional to size (PPS) sampling, with oversampling of Asian and Hispanic students.
- Treatment measure: "Have you ever been in any of the following kinds of courses or programs in high school?", where option *k* is "Career Academy"
- Treated proportion: 1,371 (8.5%) students participated, 14,826 (91.5%) who did not.

True Confounders



What covariates to include in the propensity score model

- Including variables related to exposure and the outcome in the PS model (True Confounders) decreases bias and variance.
- Including variables related to the outcome but not to exposure in the PS model does not affect bias but decreases variance.
- Including variables related to exposure but not the outcome does not affect bias but increases variance.

Identifying covariates for PS model

- Theoretical analysis of factors influencing the selection mechanism and their relationship with outcomes
- Pilot study focused on identifying the selection mechanism
- Expert reviews and interviews with participants and other persons knowledgeable about the selection process
- Use a sub-sample of the original data.

Example potential confounders of the relationship between career academy participation and future income

Variable Name in ELS Data	Variable Description
bypared	Parents' highest level of education
byhomlit	Home literacy resources
byriskfc	Number of academic risk factors in 10th grade
bystexp	How far in school student thinks will get—composite
BYS27I	Parents expect success in school
byurban	*Urbanicity of school*
byregion	*Geographic region of school*

Missing data methods for covariates

Method	Description	R Packages
Listwise deletion	Remove all observations with at least one missing value	comple.cases and na.omit functions of the stats package
Single imputation	Hot-deck imputation, regression imputation, or any method for multiple imputation can be used to obtain a single imputed data set	HotDeckImputation, hot.deck, norm, cat, mix, mice, Amelia packages

Missing data methods for covariates

Method	Description	R Packages
Multiple imputation (joint modeling)	Specify multivariate distribution for the missing data, draw imputations from posterior distribution by Markov chain Monte Carlo (MCMC) techniques	norm cat mix
Multiple imputation (multivariate imputation by chained equations)	Specify multivariate distribution -through conditional densities for each variable, draw imputations from the posterior distribution of each variable by MCMC techniques	mice

Handling missing data on covariates

- Listwise deletion: Only acceptable with very low proportion of missing data.
- Single imputation: Acceptable approach if missing data indicators are also included in the PS model.
- Multiple imputation of covariates, followed by averaging of multiple propensity scores to create single propensity score vector.
- Multiple imputation followed by separate propensity score analysis of each imputed dataset.

Estimation of Propensity Scores

- *Statistical models*: logistic regression, probit regression
  - Most widely used;
  - Rely on functional form assumption.

- *Statistical learning algorithms* (data mining methods): classification trees, boosting, bagging, random forests
  - Do not rely on functional form assumptions
  - It is not yet clear whether these methods are on average better than binomial models

Estimation of propensity scores

- Logistic regression of treatment Z on observed predictors $\mathbf{X}$
- Logistic model: $\text{logit}(Z_i = 1 | X) = \beta_0 + \beta_1 X_{1i}. + \ldots \beta_k X_{ki}$
  - Estimated *PS* : $e_i(X) = \dfrac{\exp(\text{logit}(Z_i = 1 | X))}{1 + \exp(\text{logit}(Z_i = 1 | X))}$
- The distribution of the estimated propensity scores shows how similar/different treatment and control group are.

Classification and Regression Trees

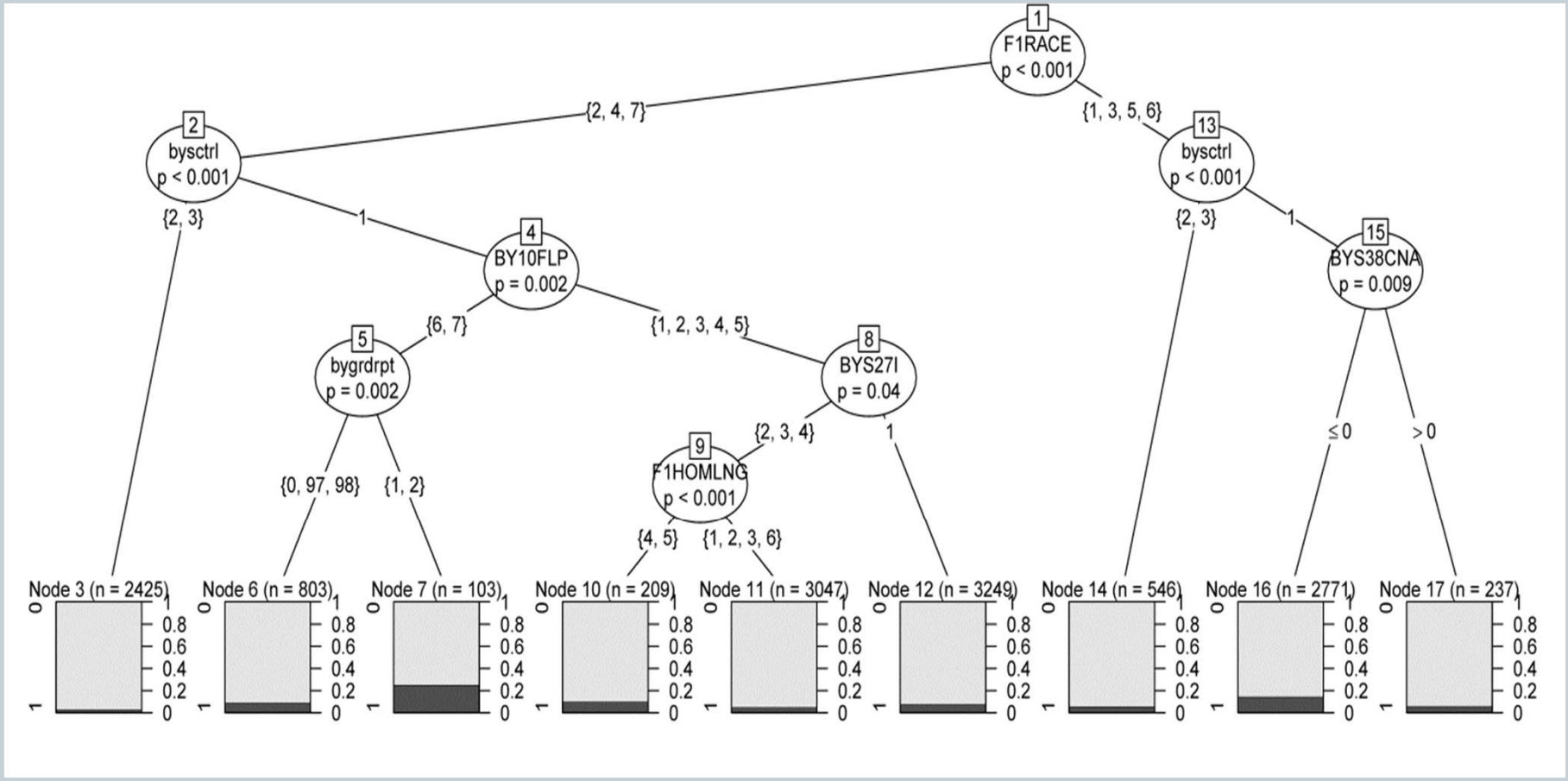
1. Each variable is split into two nodes that are more homogenous than the existing node with respect the outcome variable.

2. The classification tree algorithm calculates this gain for every possible split and selects the split resulting in largest gain in impurity reduction.

3. Variables may be used more than once to allow interactions. Trees automatically capture interactions and non-linear effects.

4. The algorithm iteratively splits variables until a stop criterion is met.

Classification tree to predict enrolling in career academy



Issues with classification trees

- The results have a high level of variability with many covariates and frequently produce poor estimates of propensity scores.
- Classification trees tend to over-fit the data, producing a tree with many branches that are due to random variation in the data and do not cross-validate to other datasets.
- To reduce over-fitting, the process of “pruning” of branches is usually employed.

Ensemble Methods

- An ensemble method builds many trees and some criterion such as best classification rate is used to create the final classification tree.
- Examples of ensemble methods:
  - **bagging**
  - **random forests**
  - Generalized boosted modeling
- Ensemble methods have been shown to work better than classification and regression trees.

Bagging and Random Forests

- Bagging runs a large number of trees with bootstrapped samples of the same size of the original sample, taken with replacement.

- Random forests is similar to bagging, except that a sample of variables is used in each iteration rather than all variables. This prevents any variable to dominate the prediction and increases variability of results.

- Steps:

1. Resample the dataset.

2. Build a prediction model on each resampled dataset.

3. Average the prediction.

Generalized Boosted Modeling

- Boosting is a general method to improve a predictor by reducing prediction error.
- GBM for propensity score estimation improves prediction of the logit of treatment assignment:

$$\text{logit}(Z_i = 1 | X)$$

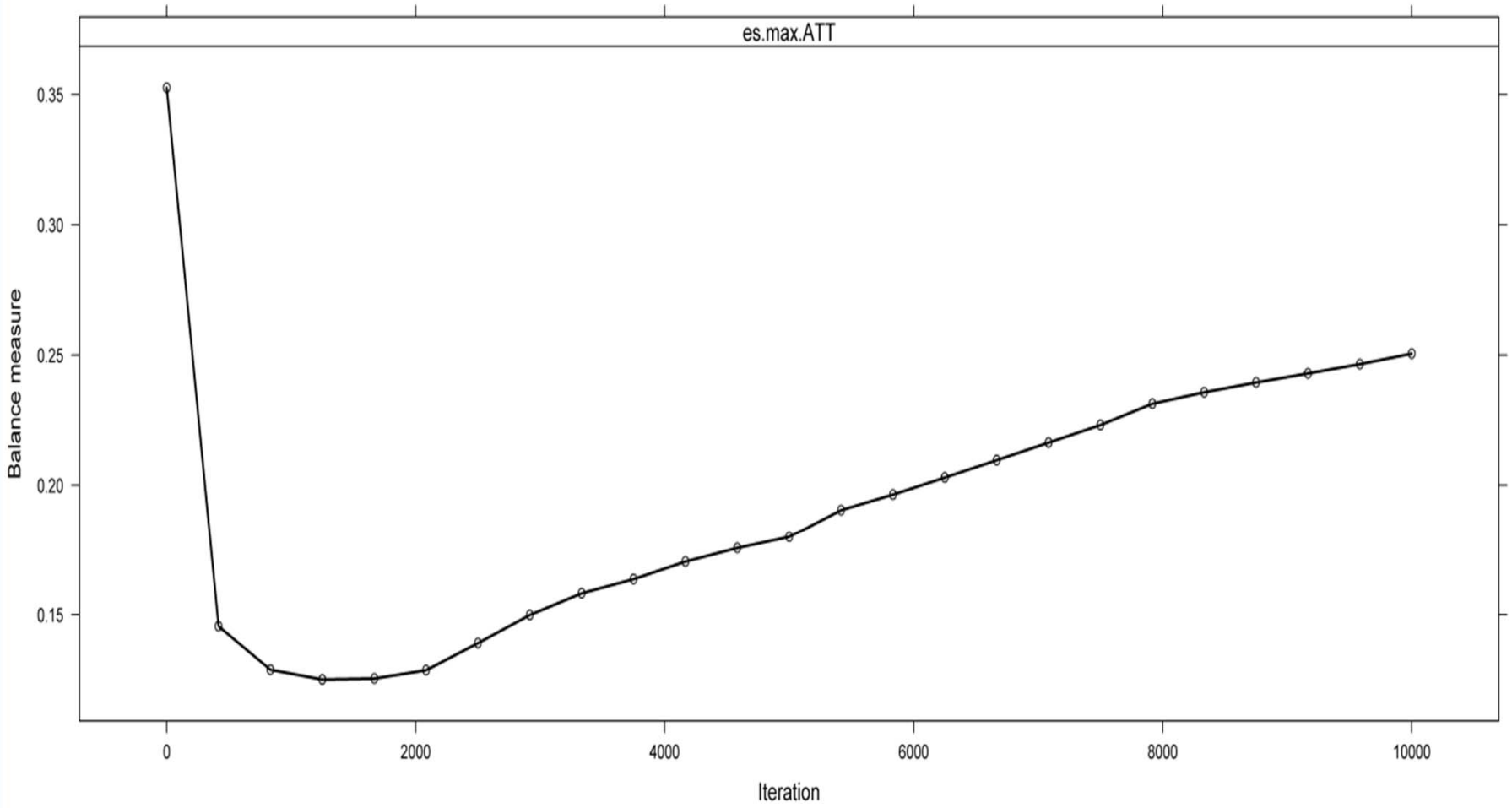
- Starting value: $\log[\bar{Z} / (1 - \bar{Z})]$
- Regression trees are used to minimize the within-node sum of squared residual:.

$$Z_i - e_i(X)$$

Stopping a GBM

- There is no defined stopping criterion, so errors decline up to a point and then increase,
- For propensity score estimation, McCaffrey et al. (2013; 2004) recommended using a measure of covariate balance, to stop the GBM algorithm the first time that a minimum covariate balance is achieved.
- There is no guarantee that better covariate balance would not be achieved if the algorithm runs additional iterations.

Determining convergence of GBM algorithm



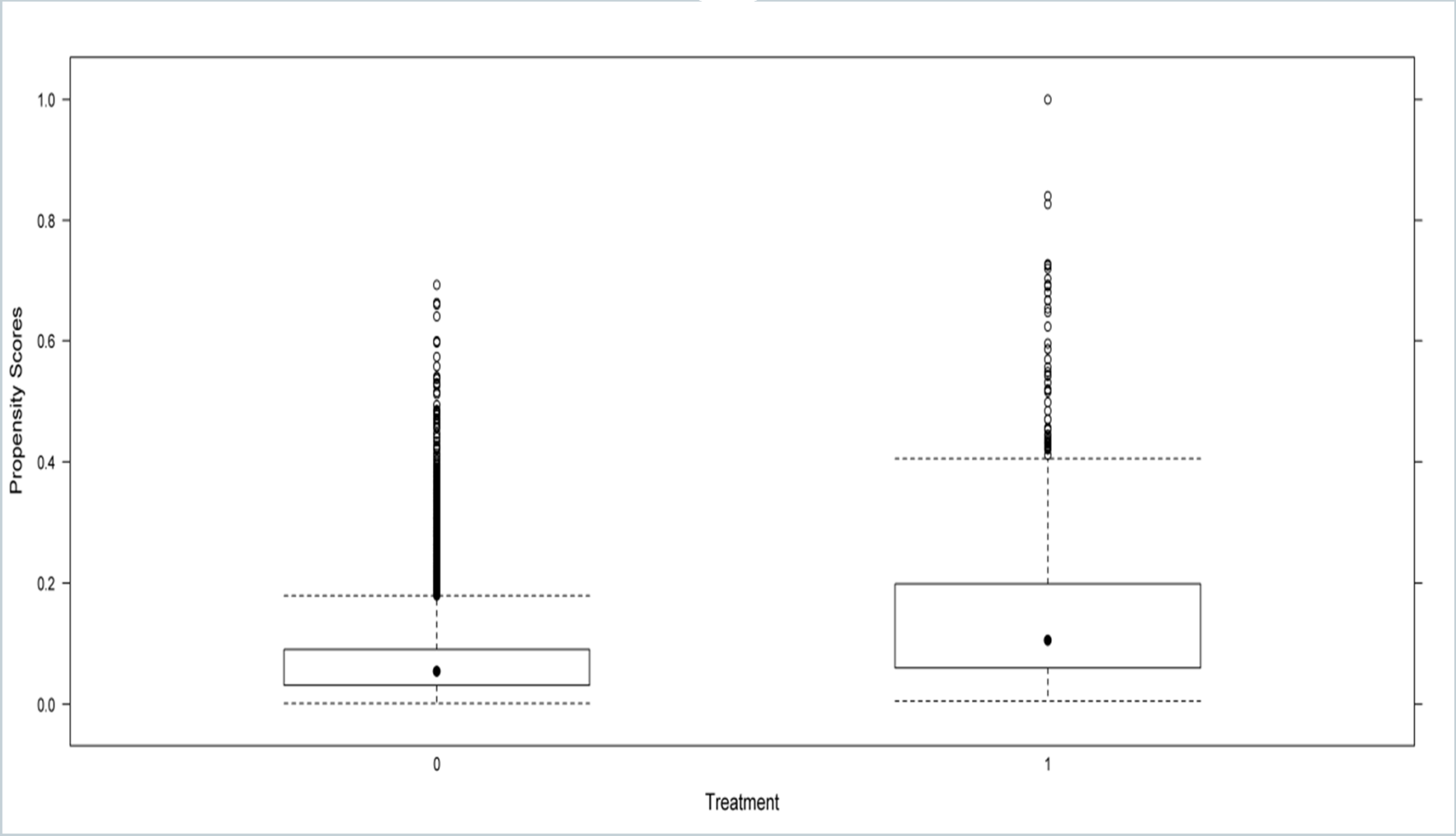
Common support

- *Common support: The area of the propensity score distribution where values exist for both groups.*
  - For *cases without common support* the treatment effect cannot be estimated
  - Lack of common support might be the reason why balance cannot be achieved
    - Balance should be checked after deletion of cases with lack of common support.
  - Discarding cases restricts the generalizability of results. However, discarding on the basis of covariate values instead of PS retains the interpretability of the treatment effect

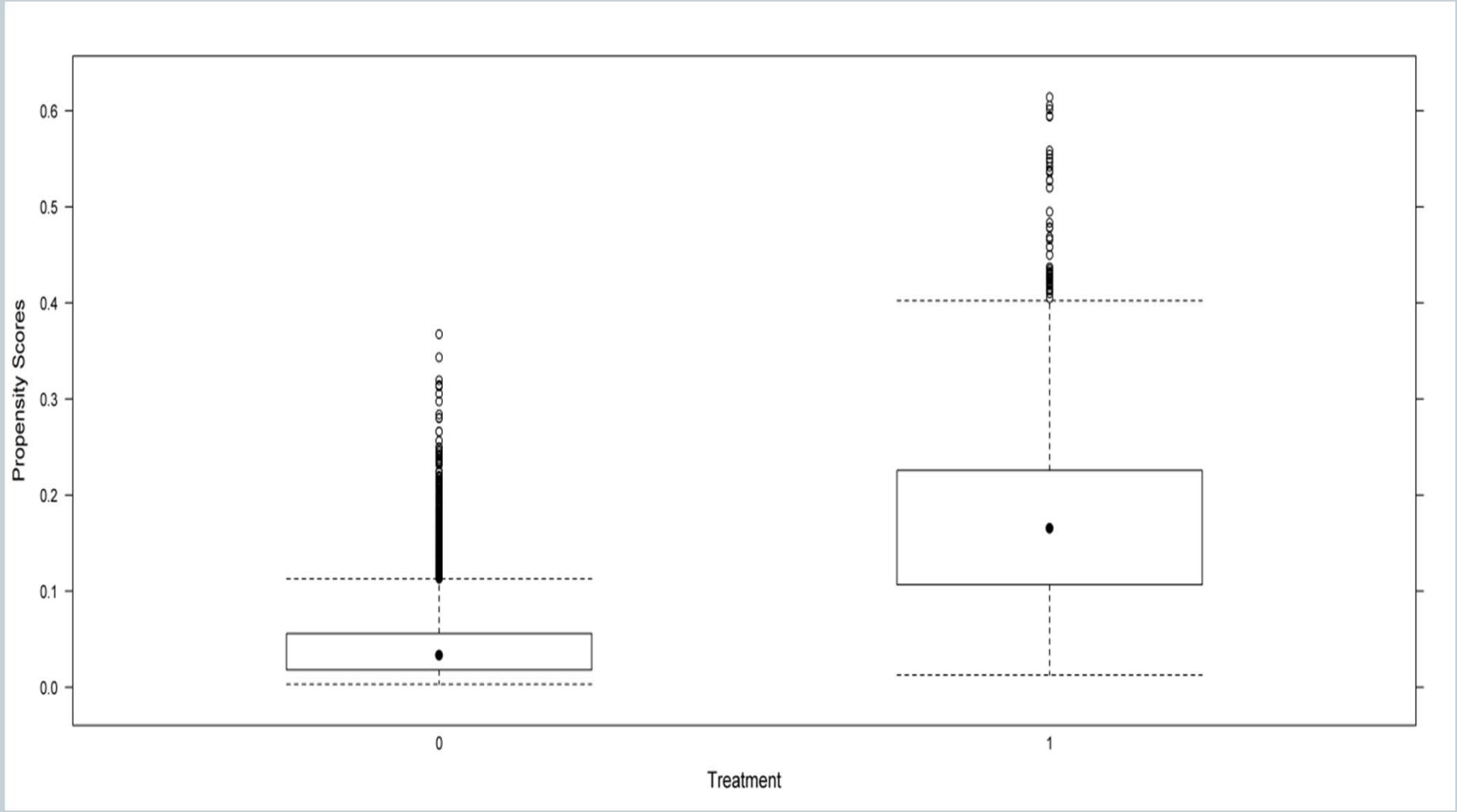
How to check common support

- Visualizations: Box-and-whisker plots and histogram
- Inspection of minimums and maximums of propensity score distributions for treated and control.
- After matching, check if there are unmatched units.
- After stratification, check if there are strata with zero cases from one of the treatment groups.

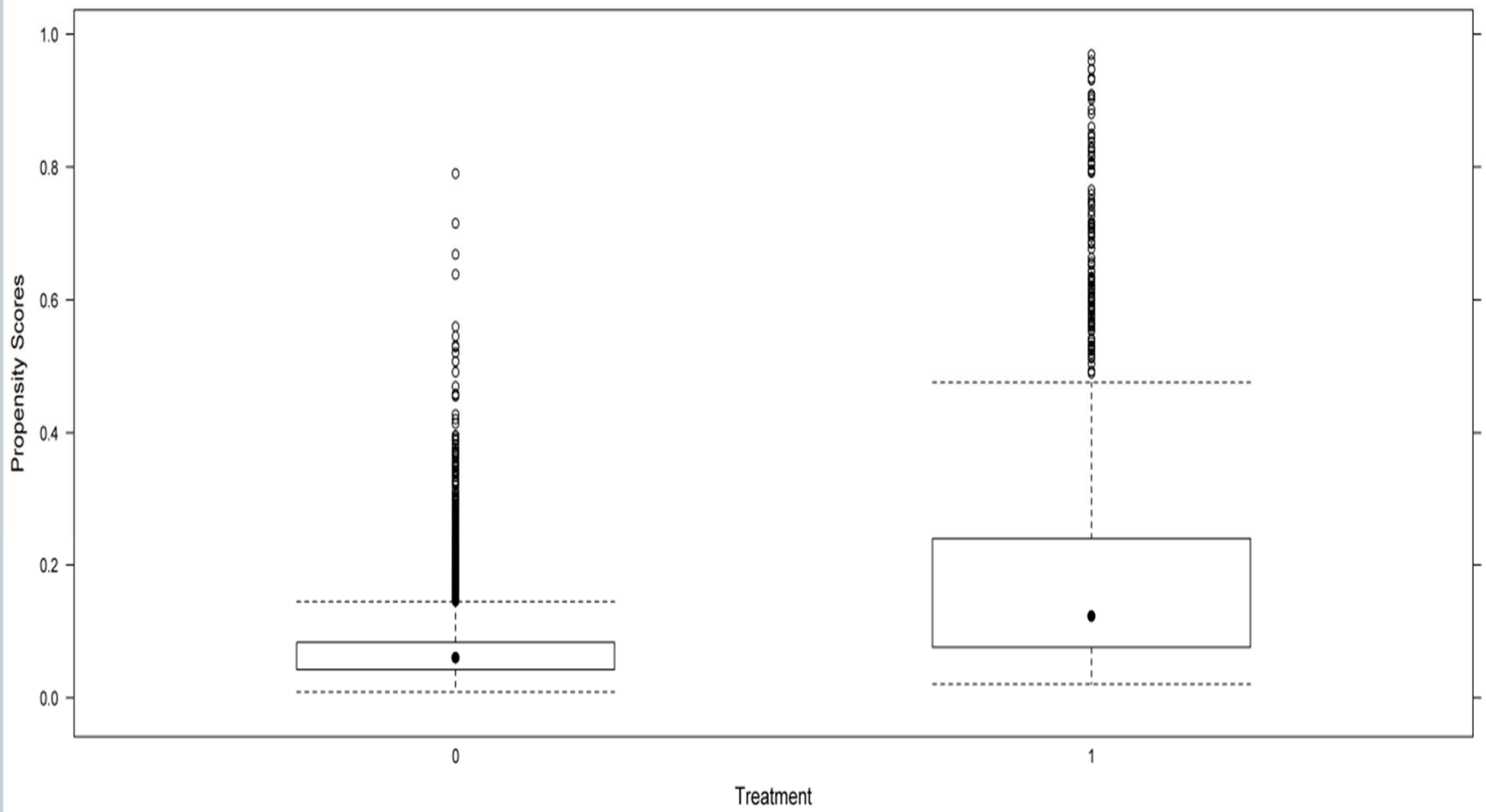
Box and Whiskers Plot for PS with Logistic Regression



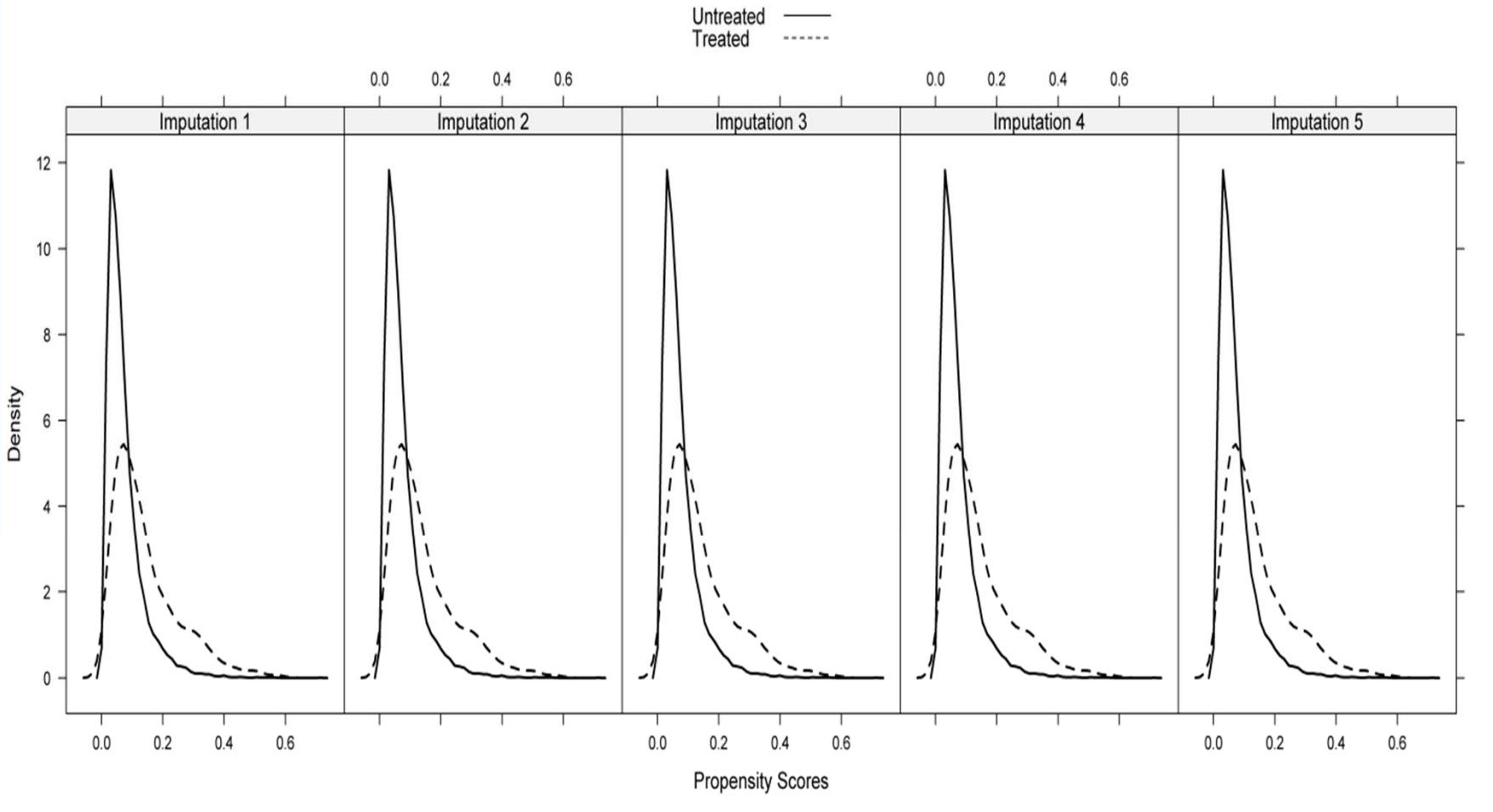
Box-and-Whiskers Plot for PS estimated with random forests



Box-and-Whiskers Plot for PS estimated with GBM



Kernel Density Plot of PS of Five Imputed Datasets with Logistic Regression



What if adequate common support and covariate balance cannot be achieved?

- Indication that groups are too *heterogeneous* for estimating a causal treatment effect (particularly if groups only overlap on their tails);
- If imbalance is not too severe one could hope that the *additional covariance adjustment* removes the residual bias due to imperfect balance;
- It is usually harder to achieve balance on a large set of covariates than on small set of covariates

R packages useful to estimate propensity scores

Package	Function	Objective
Base	glm	Fit a logistic regression for estimating propensity scores
survey	svydesign svyglm	Fit a logistic regression for estimating propensity scores with complex survey data
party	ctree	Run a classification tree to estimate propensity scores
party	cforest	Run random forests to estimate propensity scores
twang	ps	Run generalized boosted modeling to estimate propensity scores