Guidelines for Using State-of-the-Art Methods to Estimate Propensity Score and Inverse Probability of Treatment Weights When Drawing Causal Inferences

Session 2. Propensity Score Estimation

Beth Ann Griffin
Dan McCaffrey

Review of Causal Effect Estimation

- ***Assume*** **that for individuals with the same values of the covariates, the distributions of both treatment and control potential outcomes are the same for cases assigned to treatment or control**
- **Use propensity scores, probability of treatment assignment, to create treatment and control groups with similar covariates so the control group can estimate the counterfactual for the treatment group and vice versa**

Propensity Score and Balance

- Conditional on the propensity scores the distribution of covariates will be the same for treatment and control cases
- Weighting by inverse of the propensity score can undo the differential sampling of individuals with different values of covariates and make the weighted distributions of covariates the same for treatment and control
- Covariate distributions that are the same between treatment and control are *balanced*

Estimating Treatment Effects by Estimating Propensity Scores

- Propensity scores are unknown
- We need to estimate them
  - Estimation does not require modeling the outcomes
- Balance of covariates will be useful for evaluating the estimated propensity

Conventional Wisdom

- Observe all of the important Xs and include as many Xs as possible to minimize the risk of bias
- Focus on the Xs that are predictive of the outcomes and treatment assignment
  - Avoid Xs that predict treatment but not the outcome
    - Commonly call *instruments*
    - Can contribute to bias and inflate variance of treatment effect estimates
  - Including Xs that predict only predict only the outcome can help to reduce variance of treatment effect estimates

Example: Phoenix House Academy

- Does inpatient substance abuse treatment have an effect?
- Adolescent probationers in LA County assigned to substance abuse treatment at Phoenix House Academy or other group home facility
- Rich baseline data on covariates clinically selected to measures treatment needs and substance use risk
- Outcomes at 3, 6, and 12 months post intake

Balance of Pretreatment Features

Variable	treatment mean	adjusted control mean	unadjusted control mean
Treatment motivation	2.52	?	1.35
Environmental risk	30.61	?	28.94
Substance use	7.61	?	4.59
Complex behavior	12.84	?	12.11
Age	15.82	?	15.31
⋮	⋮	⋮	⋮

- The propensity score analysis should match or weight the control subjects so that their covariates are nearly the same as the treatment group's

Propensity Scores Are Intermediate Outcomes

- **Estimate propensity scores as a means to estimating treatment effects**
- **Goal is not to make the most accurate inferences about predictors of treatment but to make the most accurate inferences about the effects of treatment**

Fit Versus Balance

- **Assessing fit is important but so is achieving good balance**
  - **Best fit under common metrics might not yield best balance and best alignment of treatment and control groups**
  - **Common approach has been to build model using forward inclusion using balance of covariates as guide**
  - **Actual implementation is very much an art form**

Steps of a Common Approach

1. **Fit logistic regression model**
   - **Include pretreatment covariates that have a significant bivariate relationship with treatment assignment (possibly identified via prior research and theory or practice)**
   - **Use stepwise regression to further refine the set of pretreatment covariates**

Steps of a Common Approach

1. **Fit logistic regression model**
2. **Weight controls as $p/(1-p)$ (ATT) or treament by $1/p$ and control by $1/(1-p)$ (ATE) or stratify by p**
   - **If stratifying start with 2 or 3 strata**
   - **Eliminate control subjects that fall into strata with very few treatment subjects and for ATE eliminate strata with very few subject of one or the other group**
   - **The propensity score weights will essentially drop unmatched control subjects for ATT**

Steps of a Common Approach

1. **Fit logistic regression model**
2. **Weight controls as $p/(1-p)$ (ATT) or treament by $1/p$ and control by $1/(1-p)$ (ATE) or stratify by p**
3. **Test for differences in the mean of each covariate**
   - **The propensity score is a balancing score. There should not be any differences in covariates for subjects with roughly the same propensity score**

Steps of a Common Approach

1. **Fit logistic regression model**
2. **Weight controls as $p/(1-p)$ or stratify by p**
3. **Test for differences in the mean of each covariate**
4. **Add interactions, quadratic terms, or strata until all covariates balance**
   - **No standard strategies apply here**
   - **Our search for alternatives began when we found this step consuming much of our time**

Deciding on Good Balance

- If the propensity scores are decent, the pretreatment covariates should balance
- Checking the balance is used to determine if the propensity score model is sufficient
  - t-tests for each covariate and higher order term
  - "Standardized bias" for each covariate and higher order term

Problems with the t-test

- With many covariates we should expect some "significant" differences
- The t-statistic is not only a function of bias but also variance

$$t = \frac{\bar{x}_{\text{treat}} - \bar{x}_{\text{control}}}{\sqrt{\frac{s^2_{\text{treat}}}{n^*_{\text{treat}}} + \frac{s^2_{\text{control}}}{n^*_{\text{control}}}}}$$

- n^* is the effective sample size and is a function of the variability in the weights
- When t is small it may indicate that we have eliminated bias, or that we have weighted or so finely stratified so that we have no power

Issues with the Conventional Approach

- Variable selection (often many covariates)
- Missing data
- Interactions
- Time consuming and difficult to automate
- Testing for differences may not indicate good balance

Our Approach

- Use a *generalized boosted model* (GBM) to estimate the propensity scores
- Use balance on the covariates to guide boosting
- Compute weighted average estimate of treatment effect on the treated
- Combine covariate adjustment with weighting when estimating the treatment effect

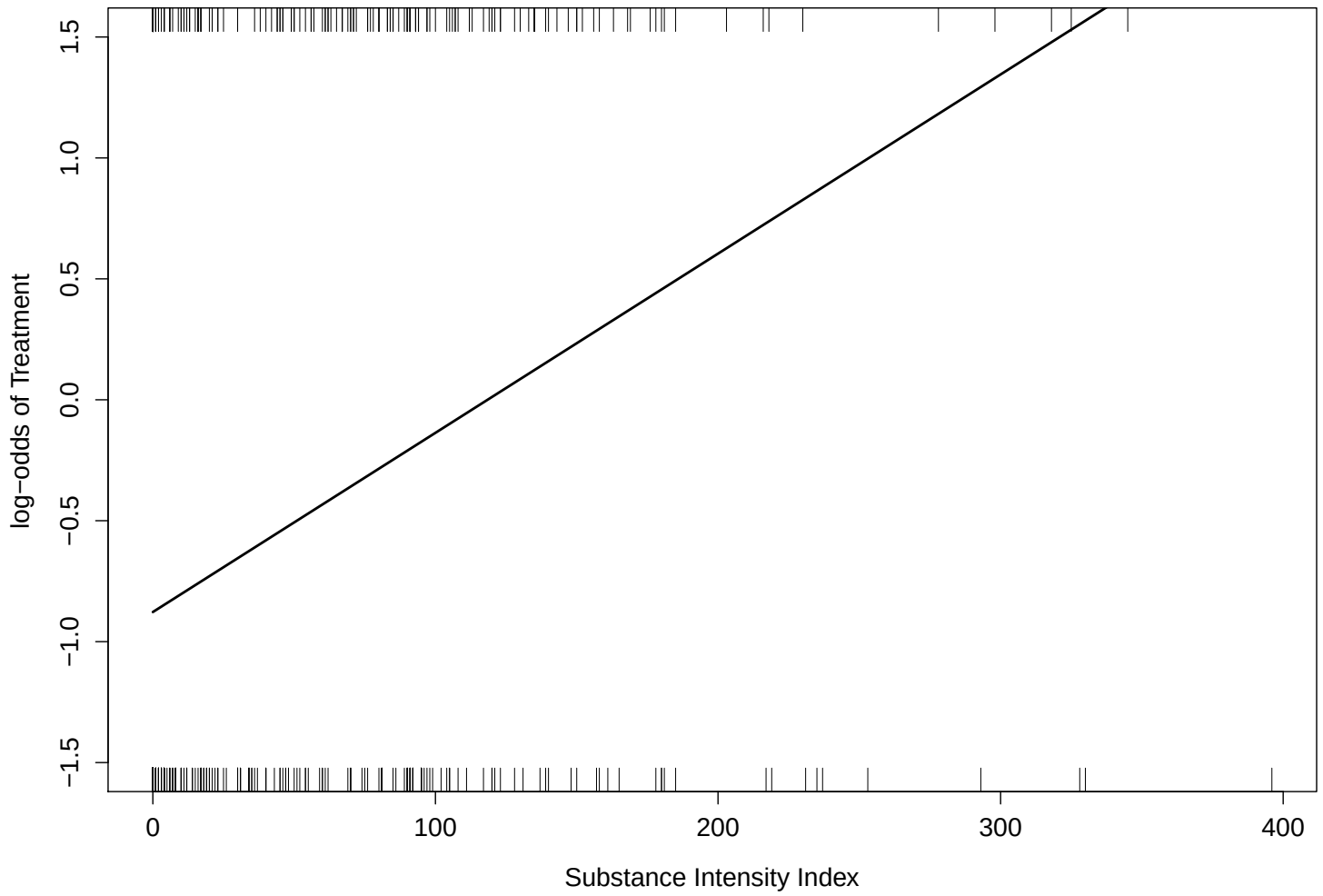
Generalized Boosting

- Data adaptive, nonparametric model function of covariates
- Combines many piecewise-constant functions of the covariates to estimate the unknown probability function
  - Use of piecewise constant functions was motivated by regression trees
- Use of piecewise constant functions means that the model can include nominal, discrete, and continuous covariates and covariates with missing values without any special instructions
- Fit is invariant to monotone transformations of covariates
- Fitting algorithm automatically selects best subset of possible piecewise functions
  - Selection of the piecewise functions automatically selects covariates for inclusion and functional form including interactions

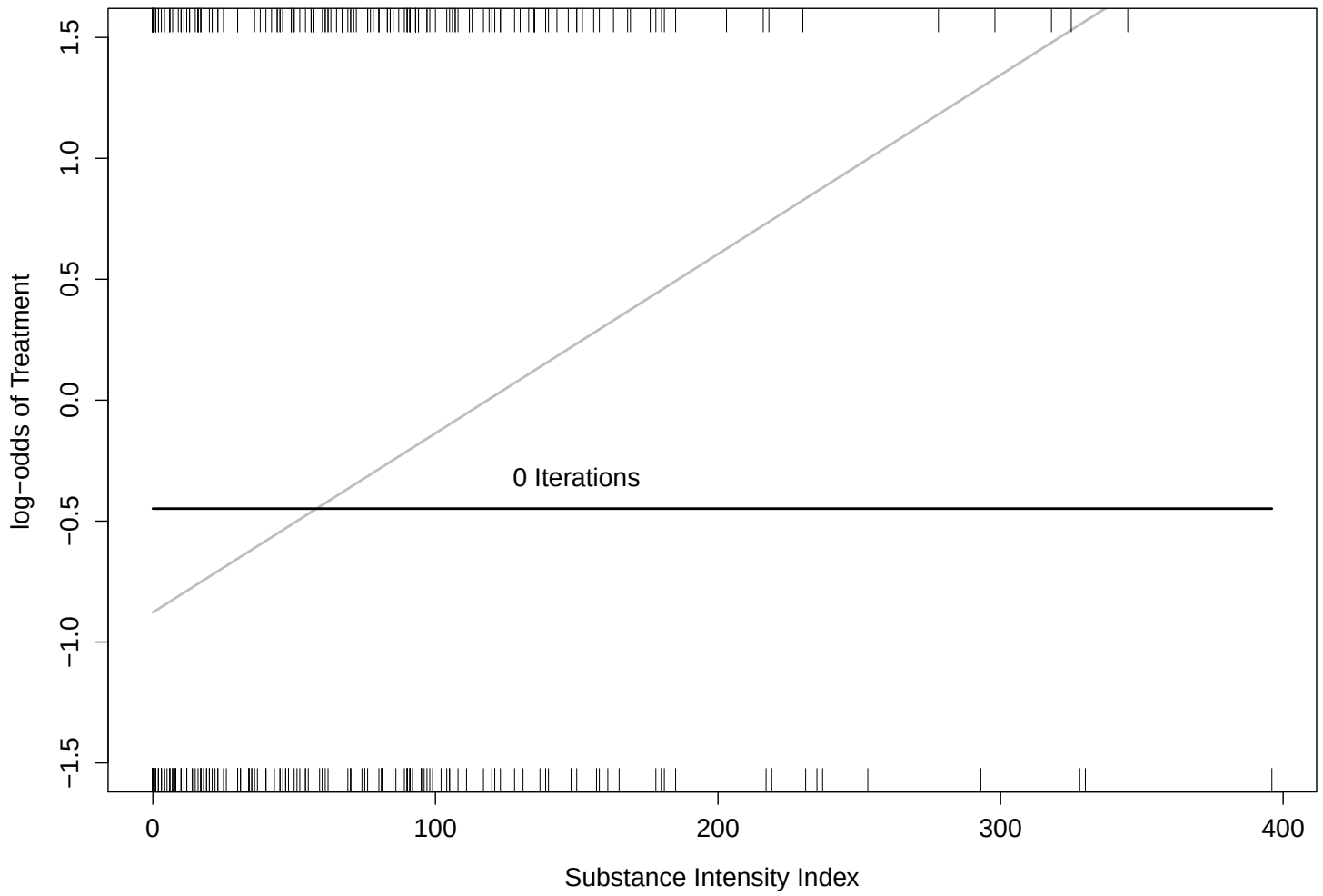
GBM Estimation

- Model fitting occurs through an automated iterative procedure
- Each iteration makes the model more complex by including more variables or making the functional form for included variables more elaborate
  - Too few iterations and the model misses important relationships between covariates and treatment assignment
  - Too many iterations and the model overfits without finding general patterns about treatment assignment
- We use balance on the covariates to determine the number of iterations
  - This is automated in our software

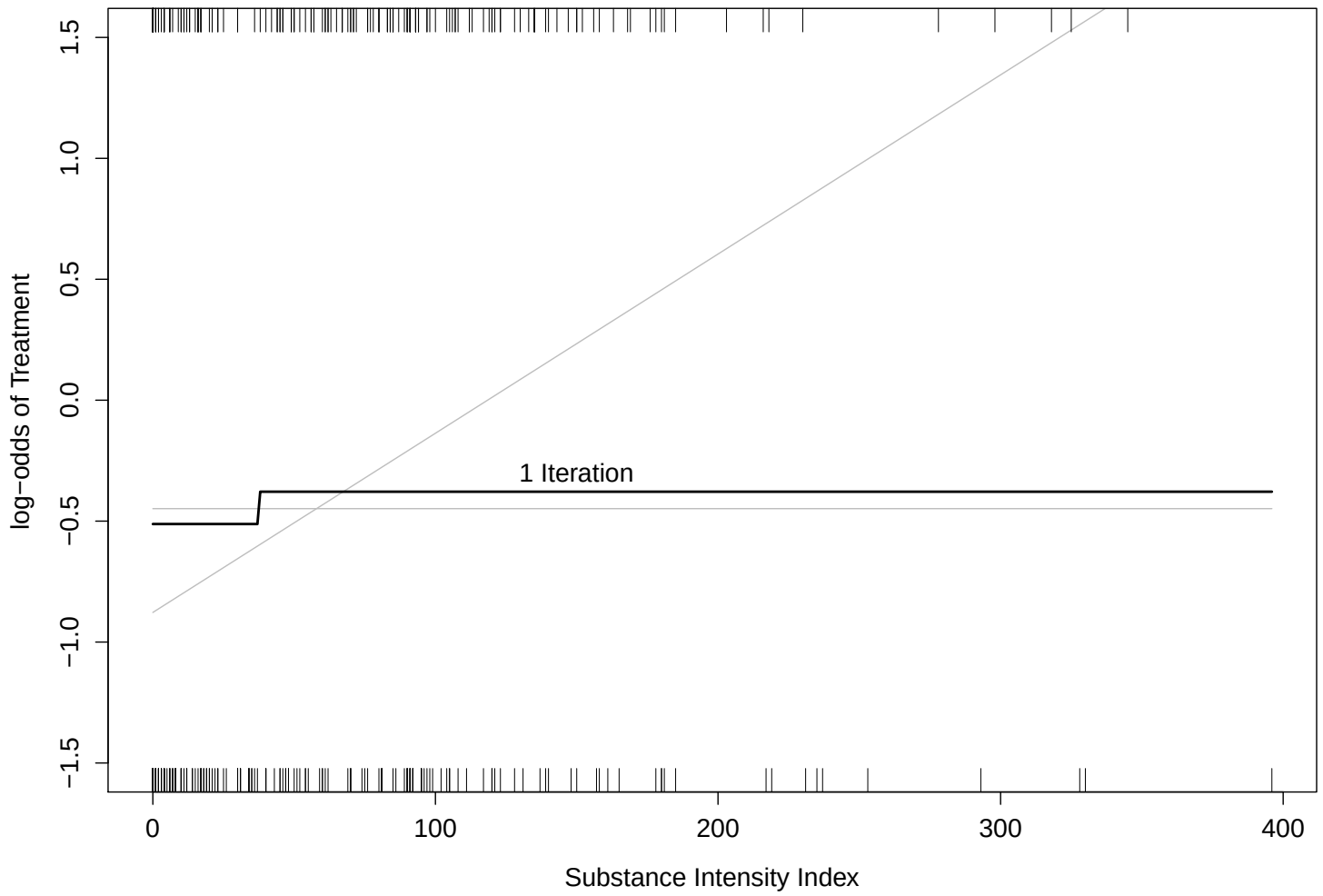
Linear Logistic Regression



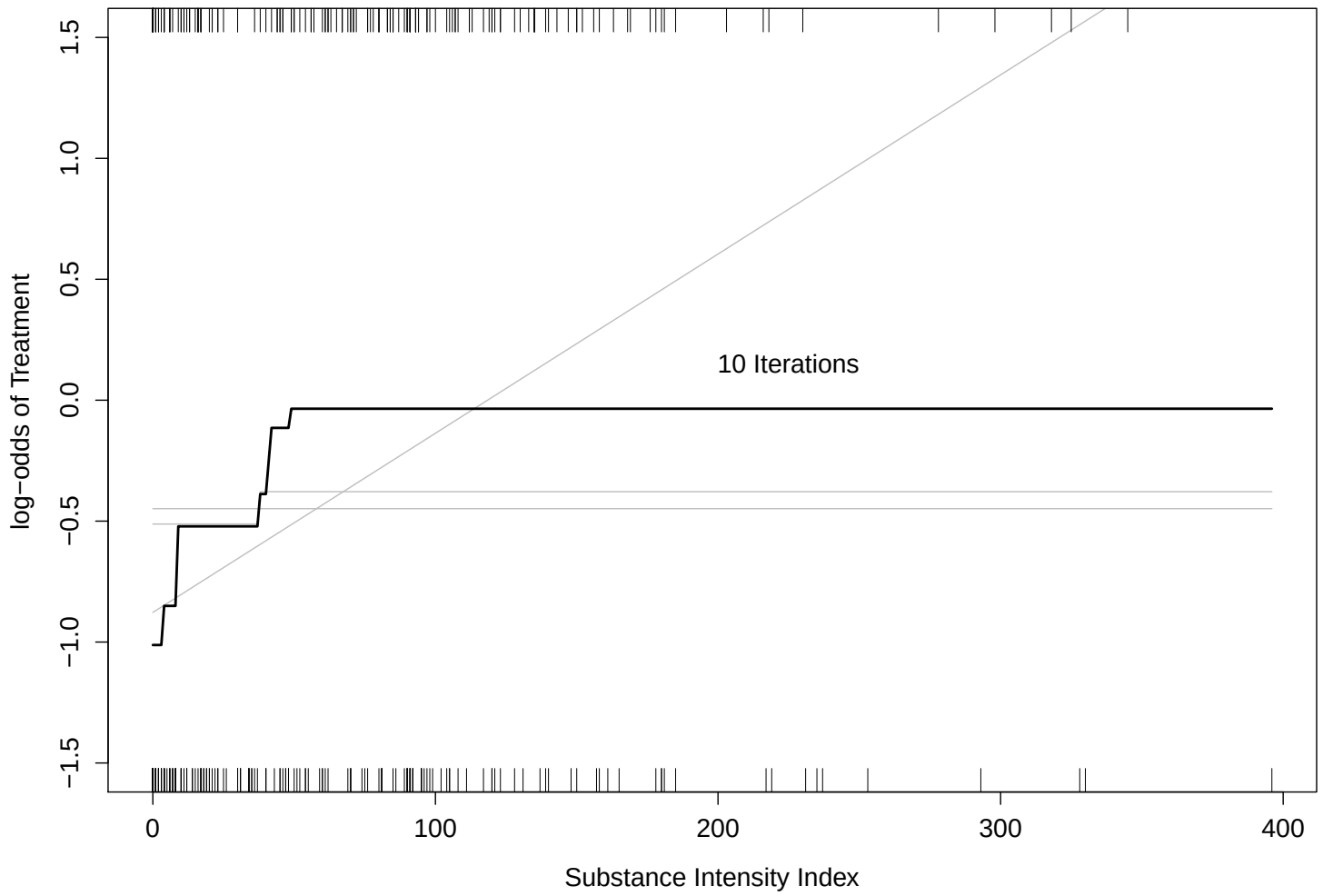
GBM Estimation



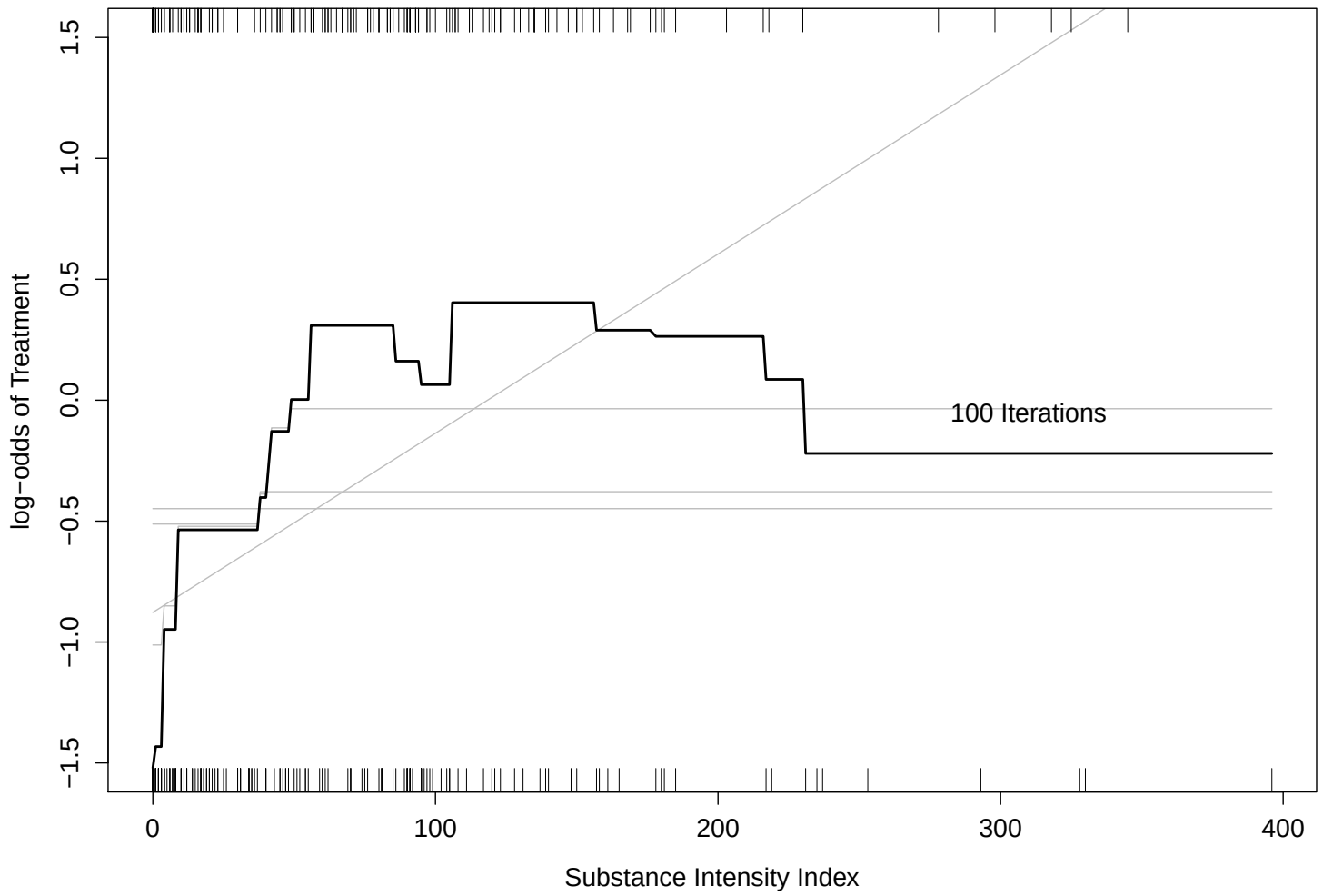
GBM Estimation



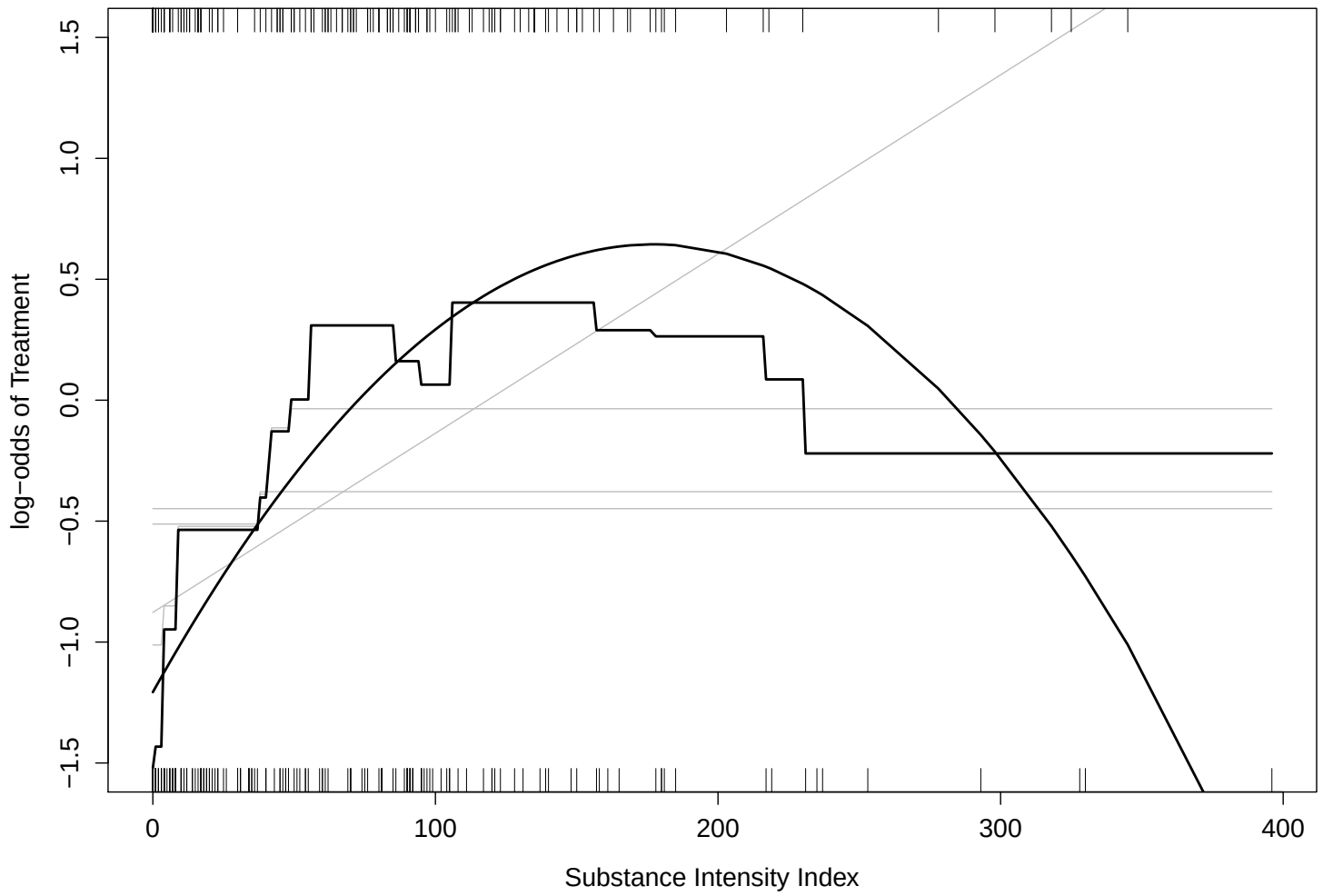
GBM Estimation



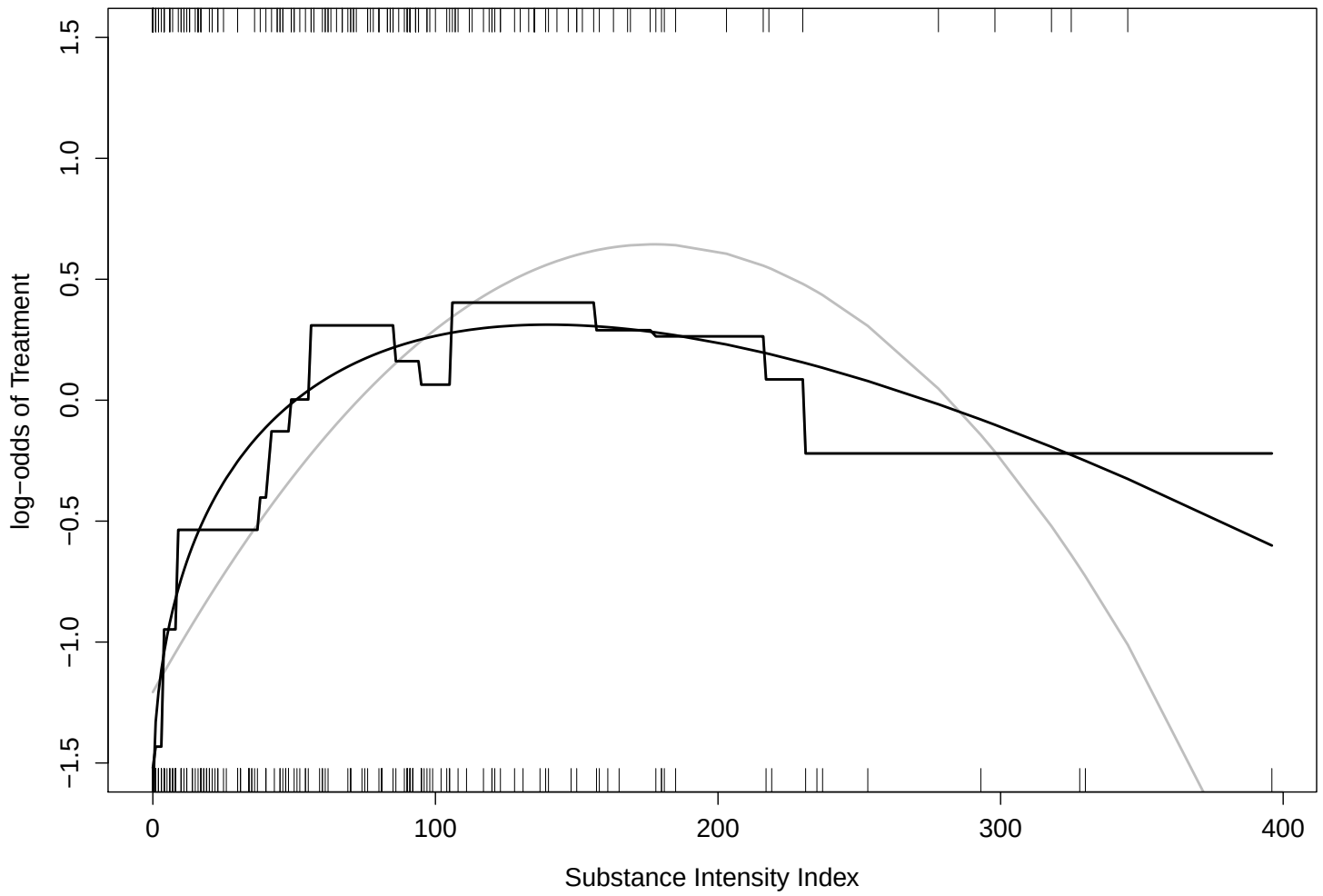
GBM Estimation



GBM Estimation



GBM Estimation



Balance of Subject Features, $C = 0$, No Iterations

Variable	treatment mean	weighted control mean	unweighted control mean	Standardized Bias weighted	Standardized Bias unweighted
Treatment motivation	2.52	2.22	1.35		0.89
Environmental risk	30.61	31.09	28.94		0.17
Substance use	7.61	6.94	4.59		0.69
Complex behavior	12.84	13.00	12.11		0.09
Age	15.82	15.76	15.31		0.56
⋮	⋮	⋮	⋮		⋮
ESS	175	107.5	274		
Average —ES—					0.307

Balance of Subject Features, 10 Iterations

		weighted	unweighted		
	treatment	control	control	effect size	
Variable	mean	mean	mean	weighted	unweighted
Treatment motivation	2.52	1.36	1.35	0.88	0.89
Environmental risk	30.61	28.97	28.94	0.17	0.17
Substance use	7.61	6.94	4.59	0.69	0.69
Complex behavior	12.84	13.00	12.11	0.08	0.09
Age	15.82	15.76	15.31	0.55	0.56
⋮	⋮	⋮	⋮	⋮	⋮
ESS	175	274.0	274		
Average —ES—				0.315	0.316

Balance of Subject Features, 100 Iterations

Variable	treatment mean	weighted control mean	unweighted control mean	effect size weighted	effect size unweighted
Treatment motivation	2.52	1.32	1.35	0.83	0.89
Environmental risk	30.61	29.24	28.94	0.14	0.17
Substance use	7.61	4.76	4.59	0.65	0.69
Complex behavior	12.84	12.33	12.11	0.06	0.09
Age	15.82	15.33	15.31	0.54	0.56
⋮	⋮	⋮	⋮	⋮	⋮
ESS	175	272.7	274		
Average —ES—				0.300	0.316

Balance of Subject Features, 1000 Iterations

Variable	treatment mean	weighted control mean	unweighted control mean	effect size weighted	effect size unweighted
Treatment motivation	2.52	1.90	1.35	0.47	0.89
Environmental risk	30.61	30.88	28.94	-0.03	0.17
Substance use	7.61	5.93	4.59	0.39	0.69
Complex behavior	12.84	13.45	12.11	-0.07	0.09
Age	15.82	15.50	15.31	0.36	0.56
⋮	⋮	⋮	⋮	⋮	⋮
ESS	175	211.1	274		
Average —ES—				0.200	.316

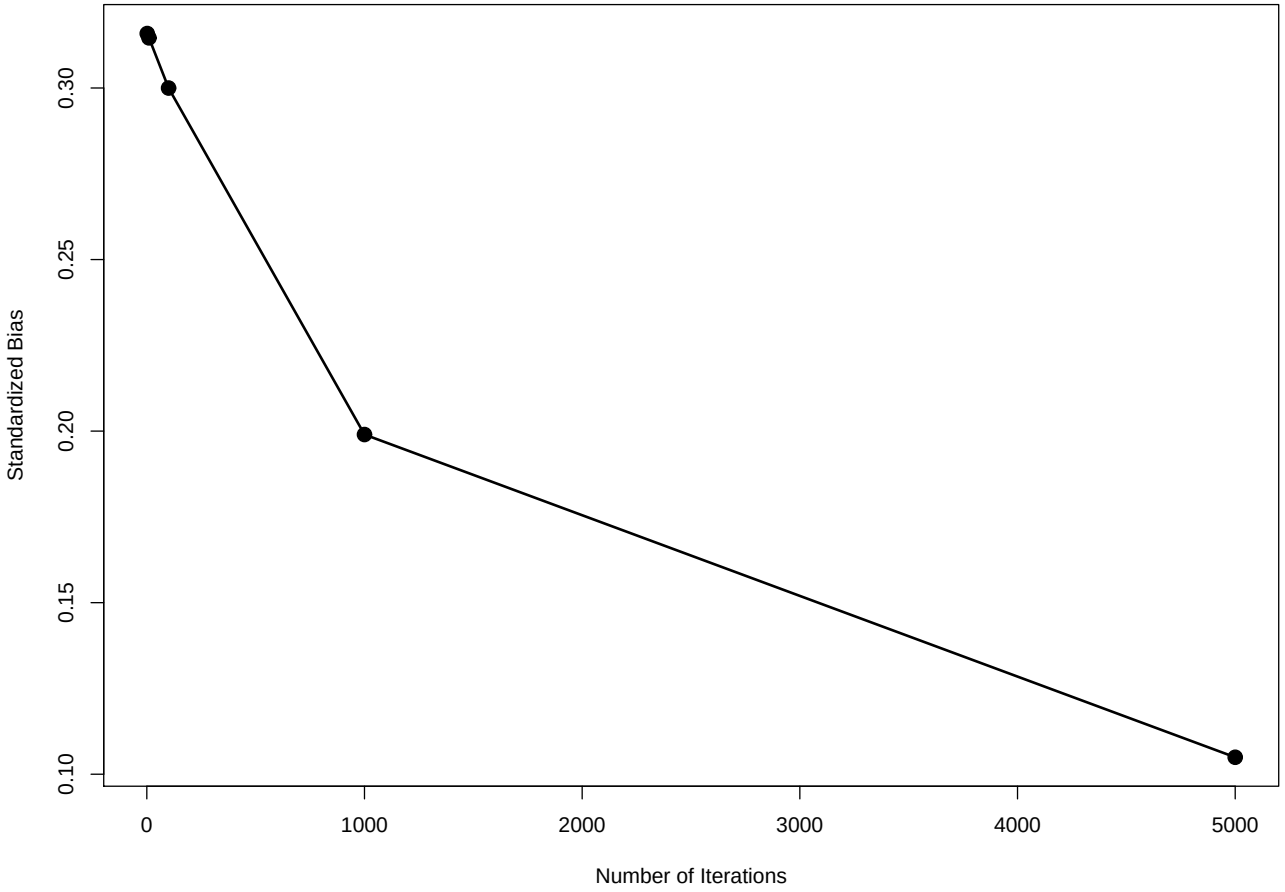
Balance of Subject Features, 5000 Iterations

Variable	treatment mean	weighted control mean	unweighted control mean	effect size weighted	effect size unweighted
Treatment motivation	2.52	2.28	1.35	0.19	0.89
Environmental risk	30.61	31.21	28.94	-0.06	0.17
Substance use	7.61	7.00	4.59	0.14	0.69
Complex behavior	12.84	13.21	12.11	-0.04	0.09
Age	15.82	15.76	15.31	0.07	0.56
⋮	⋮	⋮	⋮	⋮	⋮
ESS	175	104.3	274		
Average —ES—				0.105	0.316

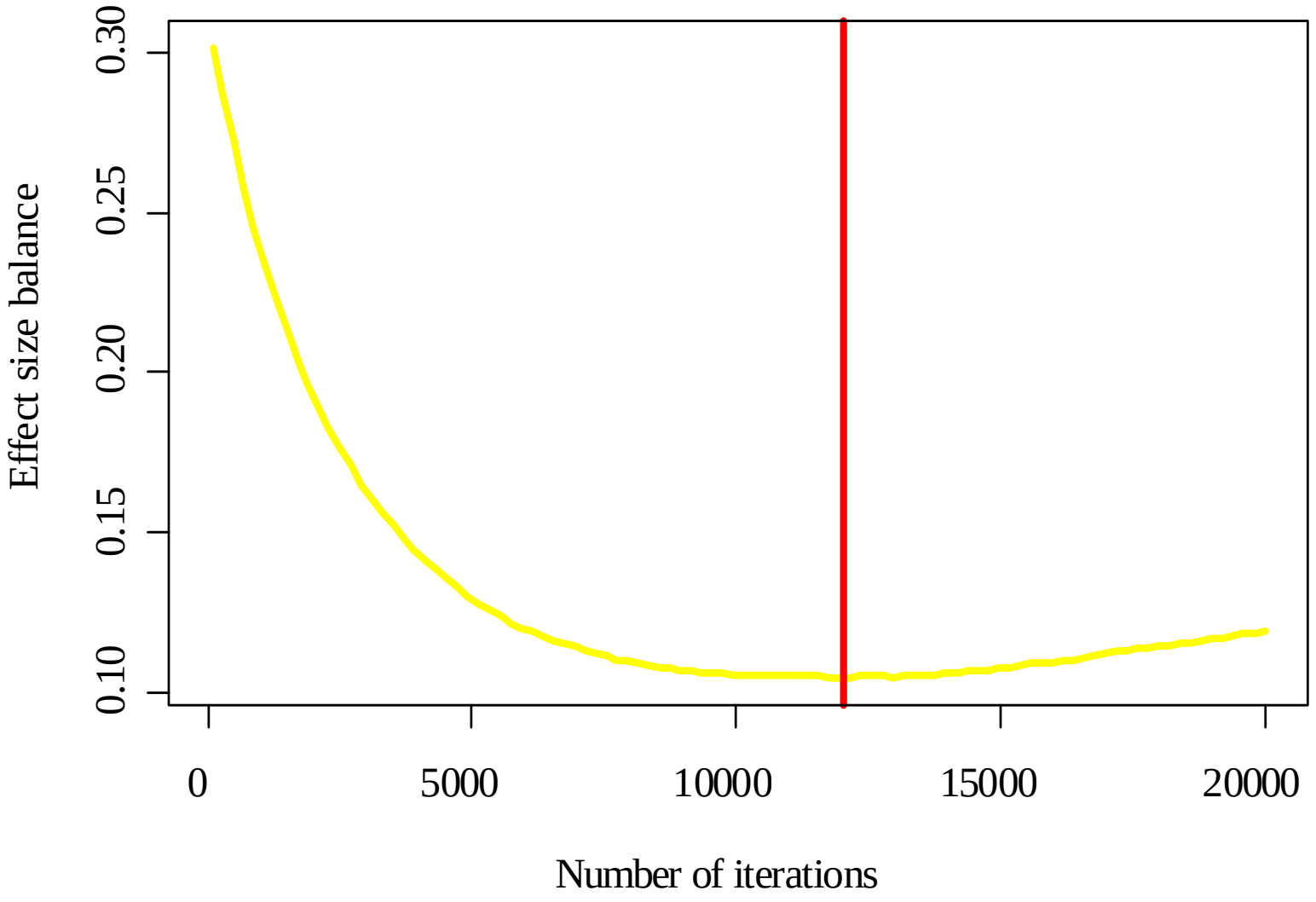
Tuning the GBM

- We must select the number of iterations
  - Within sample prediction error decreases with each additional iteration
  - Too few iterations under fit, miss features of the model, and don't achieve balance
  - Too many iterations overfit, too little variation among weights with group, and don't achieve balance
  - Somewhere in between is the best balance
- Fit with a large number of iterations
- Assess balance at each iteration
- Pick the iteration that minimizes the balance

Increasing Iterations Improves Balance



Model Selection



Assessing Balance

- Assess balance for each covariate using one of two balance metric
  - Standardized bias (or an effect size or absolute standardize mean difference)
  - Kolmogorov-Smirnov statistics
- Aggregate across covariates
  - Mean or maximum of the balance metrics for individual covariates

Standardized Bias

- A commonly used measure of balance is standardized bias

$$SB = \frac{\tilde{x}_{\text{treat}} - \tilde{x}_{\text{control}}}{s}$$

- $\tilde{x}_{\text{treat}}$ and $\tilde{x}_{\text{control}}$ are the weighted treatment and control group means
- For ATT, s equals s_{treat}, the unweighted treatment group standard deviation which involves only the treatment subjects so that manipulating the control group never affects the measure's power
- For ATE, s is the pooled within sample unweighted standard deviation
- s does not depend on the weights

Standardized Bias and Balance

- **Small value of SB indicate balance**
- **SB is on the effect size scale: Rule of thumb was 0.20 indicated good balance, now people tend to aim for less than 0.10**
- **An average of all the computed SBs can be used as a single measure of the quality of the propensity score**

Kolmogorov-Smirnov Statistic

- Standardized bias only assesses balance of the means
- If outcomes are not linearly related to covariates, then balancing the means might not be sufficient to prevent bias
- Kolmogorov-Smirnov (KS) measures the distance between two distributions and measures balance more generally than standardized bias alone
  - Let $F_T(x)$ be the empirical distribution for the treatment cases

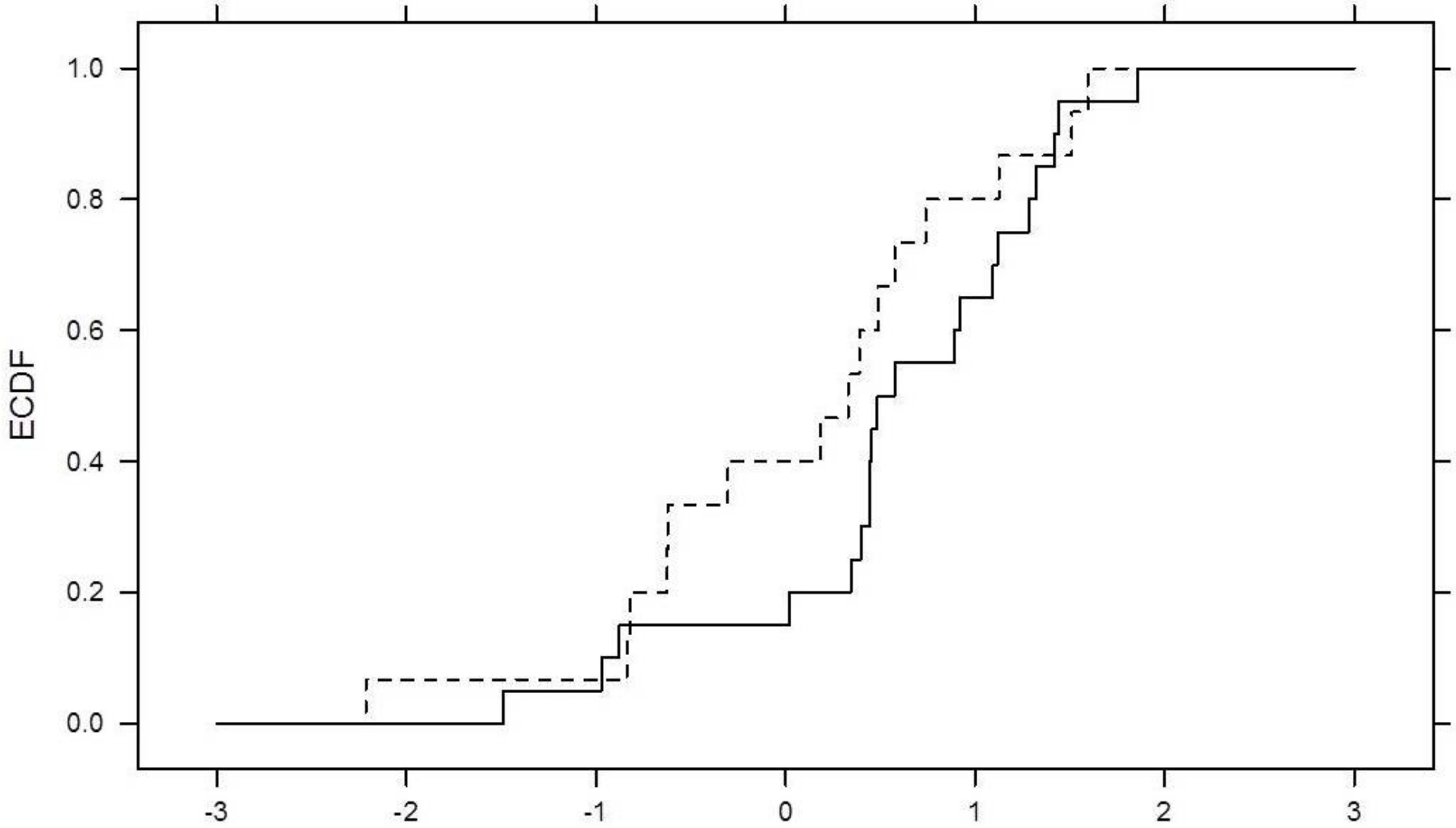
$$F_T(x) = \frac{\sum I(x_i < x)}{N_T}$$

  - Let $F_C(x)$ be the weighted empirical distribution for the control cases

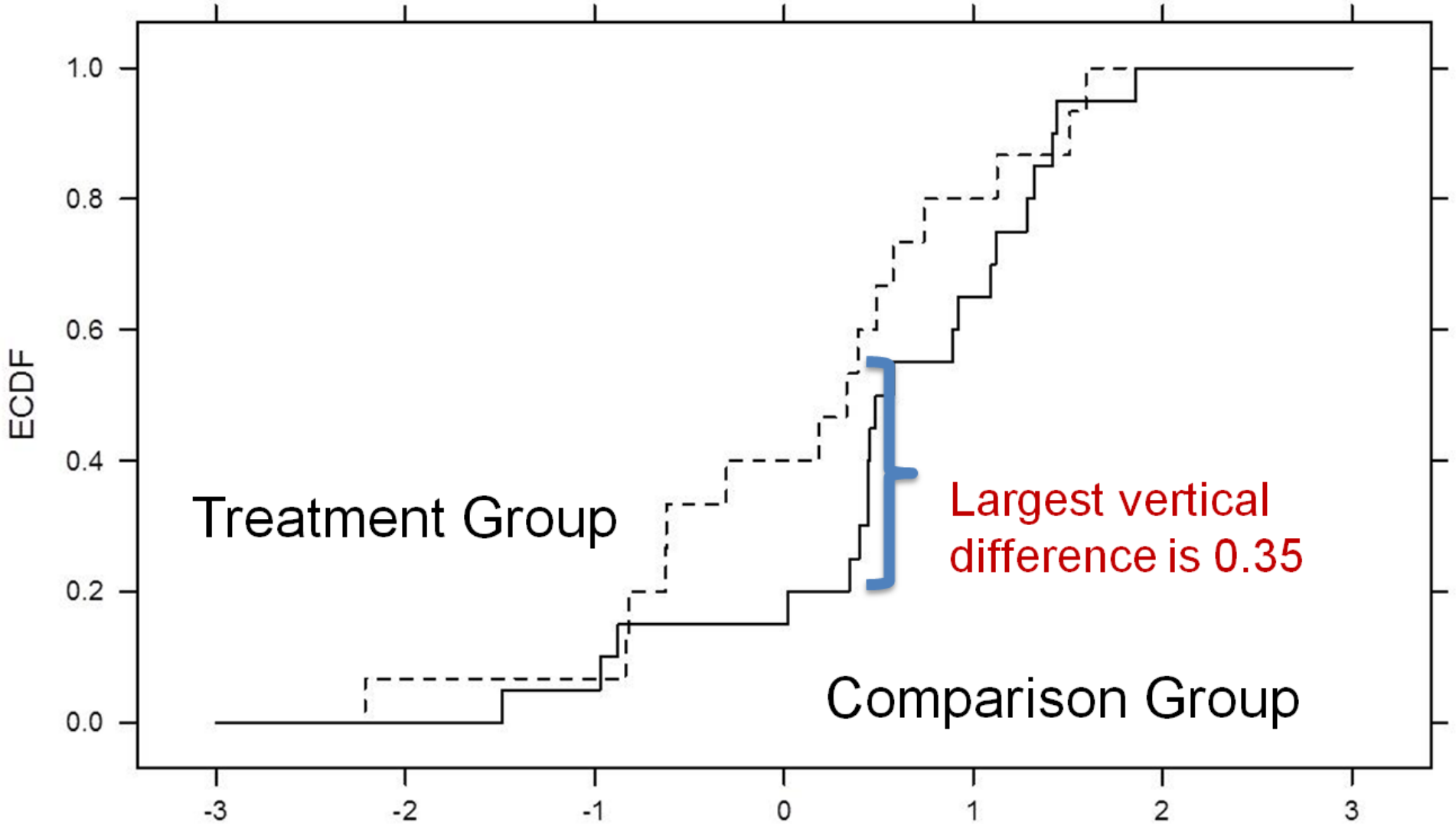
$$F_C(x) = \frac{\sum w_i I(x_i \leq x)}{\sum w_i}$$

  - $KS = max|F_T(x) - F_C(x)|$

KS Example



KS Example



Issues with KS

- There are no clear standards for what is a large or small value
- Size of KS depends on sample size
- Very useful for comparing alternative models, but less obvious if has met an objective standard for “good” balance
- No standard tests for weighted KS
- We developed approximate tests for individual covariates and the maximum across all variables
- Tests depend on the sample size

Modeling with Multiple Measures of Balance

- We consider both the KS and standardized bias
- We have a sample of each measure across multiple covariates
- We can use the maximum or the mean (or some other summary statistic) for each balance measure
- My preference is to tune GBM using multiple measures and choose the solution that appears most robust across the alternatives
- If standardized bias is much greater than .10 for any variables or somewhat greater for many variables, I consider reformulating the problem – becomes an art again

Alternative Stopping Rule Given Similar but not the Same Weights

		Standardized Bias		KS		Number of
Stop Method	ESS	maximum	mean	maximum	mean	Iterations
Mean KS Statistic	88.04	0.264	0.093	0.136	0.058	720
Mean Standardized Bias	88.39	0.261	0.093	0.134	0.058	730
Maximum KS Statistic	95.01	0.269	0.099	0.132	0.060	572
Maximum Standardized Bias	88.49	0.260	0.094	0.135	0.058	742

Estimating the Treatment Effect

- **After balancing the only *observed* difference between the two groups is the treatment assignment**
- **Average Treatment Effect on the Treated (ATT) estimate:**

$$\widehat{TE} = \frac{\sum_{i \in T} y_i}{N_T} - \frac{\sum_{i \in C} w_i y_i}{\sum_{i \in C} w_i}$$

- **Average Treatment Effect (ATE) estimate:**

$$\widehat{TE} = \frac{\sum_{i \in T} w_i y_i}{\sum_{i \in T} w_i} - \frac{\sum_{i \in C} w_i y_i}{\sum_{i \in C} w_i}$$

Variable Weights Can Add To Variance in Estimated Treatment Effect

- $Y_t = E[Y_t \mid X] + e, \text{ for t} = 0, 1$

$$
\begin{aligned}
TE &= \frac{\sum i \in T w_i y_i}{\sum i \in T w_i} - \frac{\sum i \in C w_i Y_i}{\sum i \in C w_i} \\
&= \left(\frac{\sum i \in T w_i E[Y_i \mid X_i]}{\sum i \in T w_i} - \frac{\sum i \in C w_i E[Y_i \mid X_i]}{\sum i \in C w_i} \right) + \\
&\qquad \left(\frac{\sum i \in T w_i e_i}{\sum i \in T w_i} - \frac{\sum i \in C w_i e_i}{\sum i \in C w_i} \right)
\end{aligned}
$$

Variable Weights Can Add To Variance in Estimated Treatment Effect

- $Y_t = E[Y_t \mid X] + e, \text{ for t} = 0, 1$

$$
\begin{aligned}
TE &= \frac{\sum i \in T w_i Y_i}{\sum i \in T w_i} - \frac{\sum i \in C w_i Y_i}{\sum i \in C w_i} \\
&= \left(\frac{\sum i \in T w_i E[Y_i \mid X_i]}{\sum i \in T w_i} - \frac{\sum i \in C w_i E[Y_i \mid X_i]}{\sum i \in C w_i} \right) + \\
&\quad \left(\frac{\sum i \in T w_i e_i}{\sum i \in T w_i} - \frac{\sum i \in C w_i e_i}{\sum i \in C w_i} \right)
\end{aligned}
$$

Variable Weights Can Add To Variance in Estimated Treatment Effect

- $Y_t = E[Y_t \mid X] + e, \text{ for t} = 0, 1$

$$
\begin{aligned}
TE &= \frac{\sum i \in T w_i Y_i}{\sum i \in T w_i} - \frac{\sum i \in C w_i Y_i}{\sum i \in C w_i} \\
&= \left(\frac{\sum i \in T w_i E[Y_i \mid X_i]}{\sum i \in T w_i} - \frac{\sum i \in C w_i E[Y_i \mid X_i]}{\sum i \in C w_i} \right) + \\
&\qquad \left(\frac{\sum i \in T w_i e_i}{\sum i \in T w_i} - \frac{\sum i \in C w_i e_i}{\sum i \in C w_i} \right)
\end{aligned}
$$

Variable Weights Can Add To Variance in Estimated Treatment Effect

$$\text{var}\left(\frac{\sum_{i\in T} w_i e_i}{\sum_{i\in T} w_i} - \frac{\sum_{i\in C} w_i e_i}{\sum_{i\in T} w_i}\right) \approx \text{var}(e_i)\frac{\sum_{i\in T} w_i^2}{\left(\sum_{i\in T} w_i\right)^2} - \text{var}(e_i)\frac{\sum_{i\in C} w_i^2}{\left(\sum_{i\in C} w_i\right)^2}$$

- $\frac{\sum i\in T w_i^2}{(\sum i\in T w_i)^2} \geq \frac{1}{N_1}$
  - **Same for control**
- **Variable weights can increase the variance**

Effective Sample Size

- The effective sample size (ESS) is the number of observations in the control group that "match" with the treatment group
- ESS is the number of independent observations from a simple random sample that would yield the same precision as the N_c weighted observations

$$ESS = \frac{\left(\sum_{i \in C} w_i\right)^2}{\sum_{i \in C} w_i^2}$$

Advantages of GBM

1. **Excellent estimation of $p(X)$**
2. **Balances the Xs with little effort**
3. **The resulting model handles continuous, nominal, ordinal, and missing Xs**
4. **Invariant to 1-to-1 transformations of the Xs**
5. **Model higher interaction terms with more complex regression trees**
6. **Implemented in R in the `twang` library with many tools to make propensity score estimation easy**

