FINDING SUBGROUPS OF ENHANCED TREATMENT EFFECT

Jeremy M G Taylor

Jared Foster

University of Michigan

Steve Ruberg

Eli Lilly

1. INTRODUCTION and MOTIVATION
2. PROPOSED METHOD
    - Random Forests
    - Classification and Regression Trees
3. SIMULATED DATA

SETTING: RANDOMIZED CLINICAL TRIAL

- Two treatment groups
- Binary outcome
  - Efficacy
  - Toxicity
- Lots of baseline covariates
  - Range 5 to 100
- Trial is already completed, small or marginal overall effect

GOAL

- Find subgroup of patients with enhanced treatment effect, if it exists
- Issues
  - What do you mean by enhanced?
  - Desire subgroup to be based on a small number of covariates
  - What is the strategy for finding the subgroup
  - Can you provide honest estimates of how good the subgroup is.

SEARCHING FOR SUBGROUPS IN RANDOMIZED CLINICAL TRIAL DATA IS A STATISTICAL NO-NO

- Data dredging
- Mining the data
- Overfitting the data
- Look hard enough you will find something
- Sample sizes tend not to be large enough to find subgroups

Large literature on dangers of subgroup analysis

- Pocock et al 2002
- Rothwell et al 2005
- Lagakos 2009
- Brookes et al 2001, 2004
- Cui et al 2002
- Yusuf et al 1991
- Assman et al 2000
- Examples of people finding sign of the zodaic being important (Peto et al 1995)
- Message: use extreme caution in interpreting subgroups

Consensus opinion: Need a predefined plan for subgroup analysis

- Interpretation 1. Predefine the subgroups you are going to look at
- Interpretation 2. Predefine the strategy for searching for subgroups

A Clinicostatistical Tragedy (Feinstein 1998)

- Believes there is a patho-physiology reason for existence of categories
- Believes statistical doctrines have become too dominant
- ”Potential tragedy now is what may seem to be good statistics will be bad science”

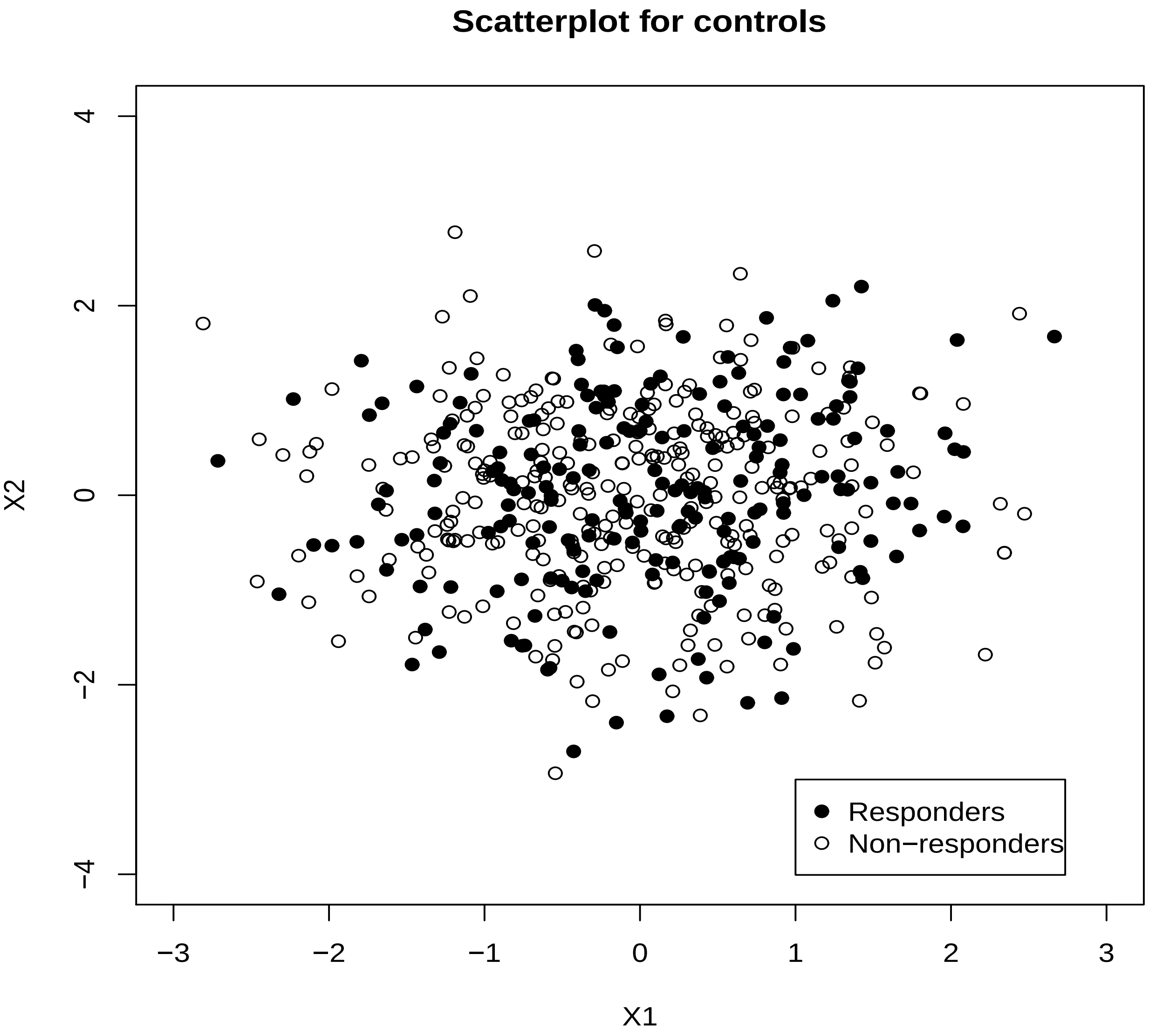
Need for validation

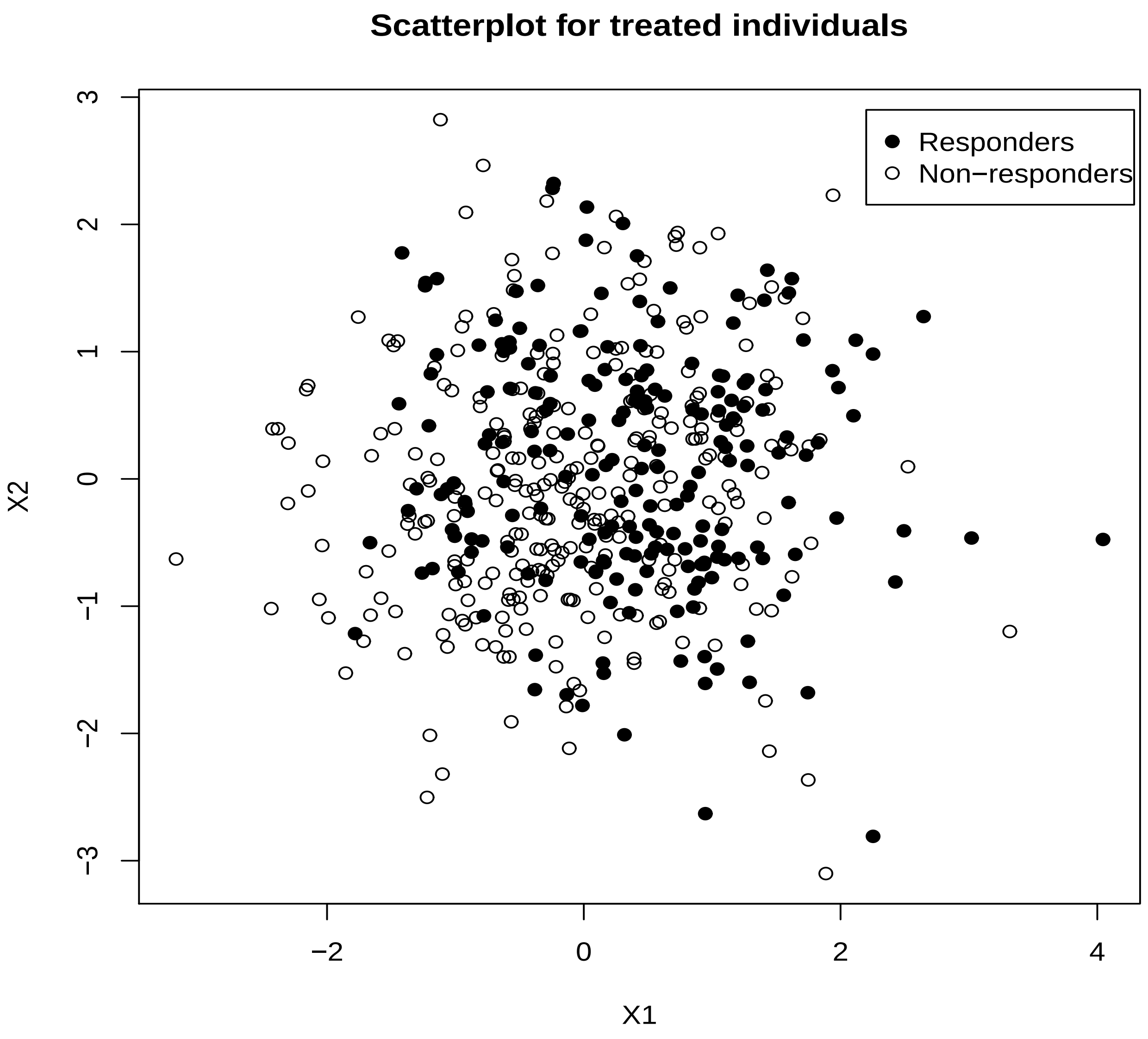
- External validation: Ideally find subgroup in one trial, validate in other trials
- Internal validation: Try to give honest estimate of quality of subgroup using the same dataset

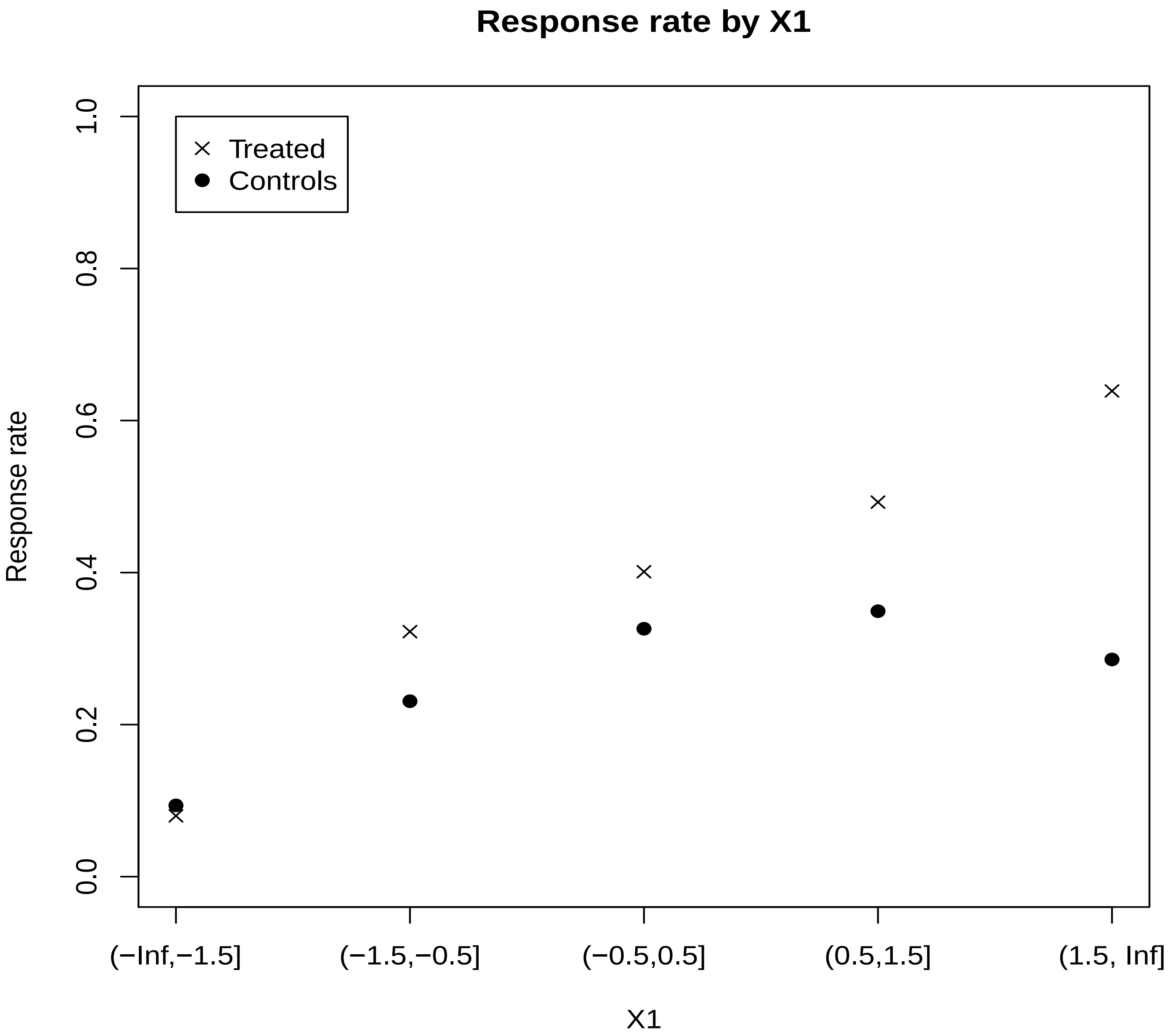
Notation

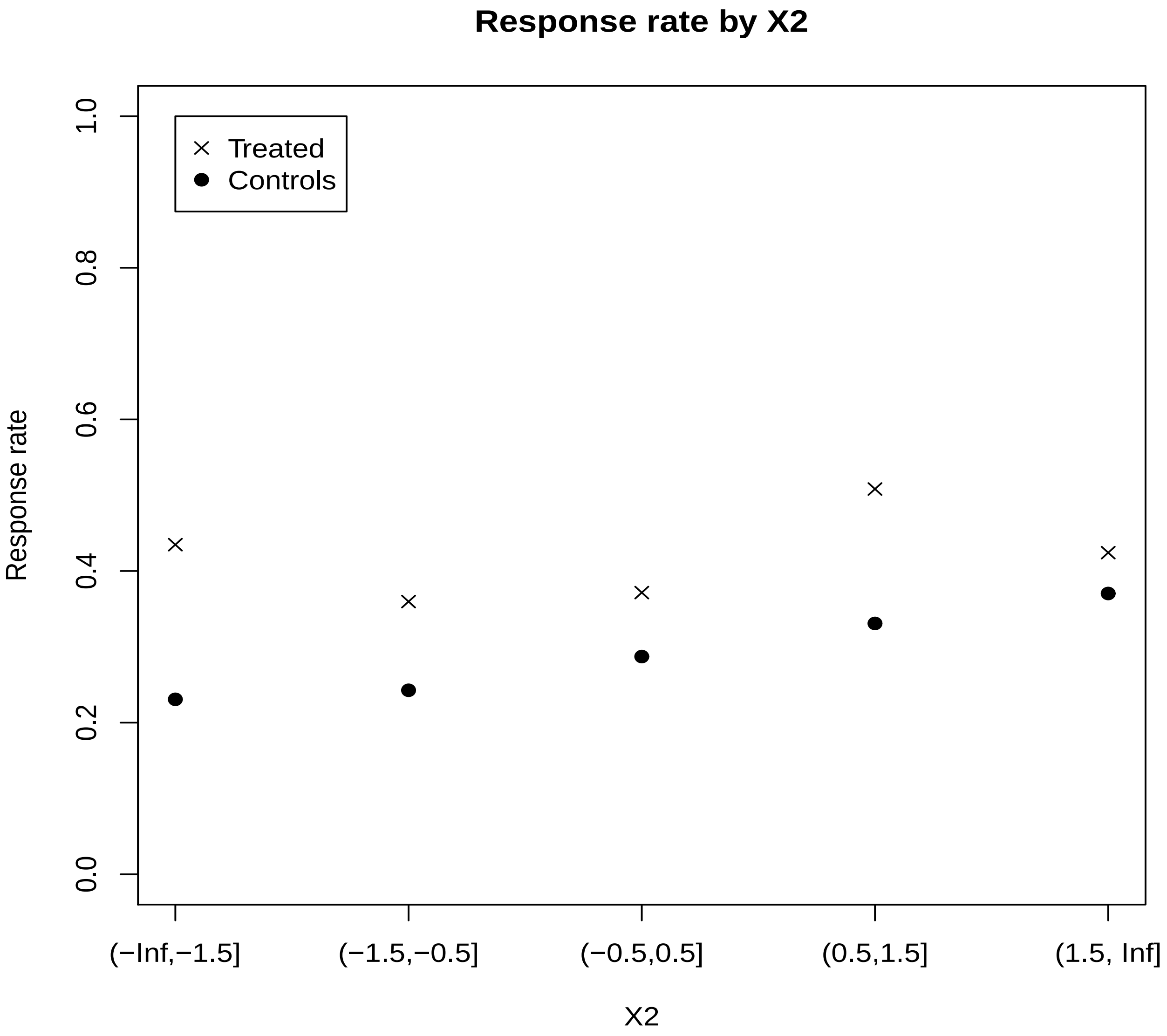
- Y = outcome, binary
- T = treatment group, binary
- $X_1, ..., X_p$ baseline covariates
- $P(Y = 1|T, X)$
- A subgroup (A) is a region of the design space
  - eg $A = \{X_1 > 3\}$
  - eg $A = \{X_1 > 3 \text{ and } X_7 < 6)\}$
  - eg $A = \{2X_1 + 3X_4 < 2\}$

- Artificial data
  - 1000 observations
  - 15 X's
  - Generated from $logit(P(Y = 1)) = -1 + 0.5^*X_1 + 0.5^*X_2 - 0.5^*X_7 + 0.1^*T + 0.5^*X_2^*X_7 + 0.95^*TI(X \in A)$
  - $A = \{X_1 > 0, X_2 < 0\}$, 25% of observations
  - Treatment group response rate = 0.408
  - Control group response rate = 0.290


Scatterplot for controls
X1
X2
−3
−2
−1
0
1
2
3
−4
−2
0
2
4
● Responders
○ Non−responders


Scatterplot for treated individuals
● Responders
○ Non−responders
X1
X2
−2
0
2
4
−3
−2
−1
0
1
2
3

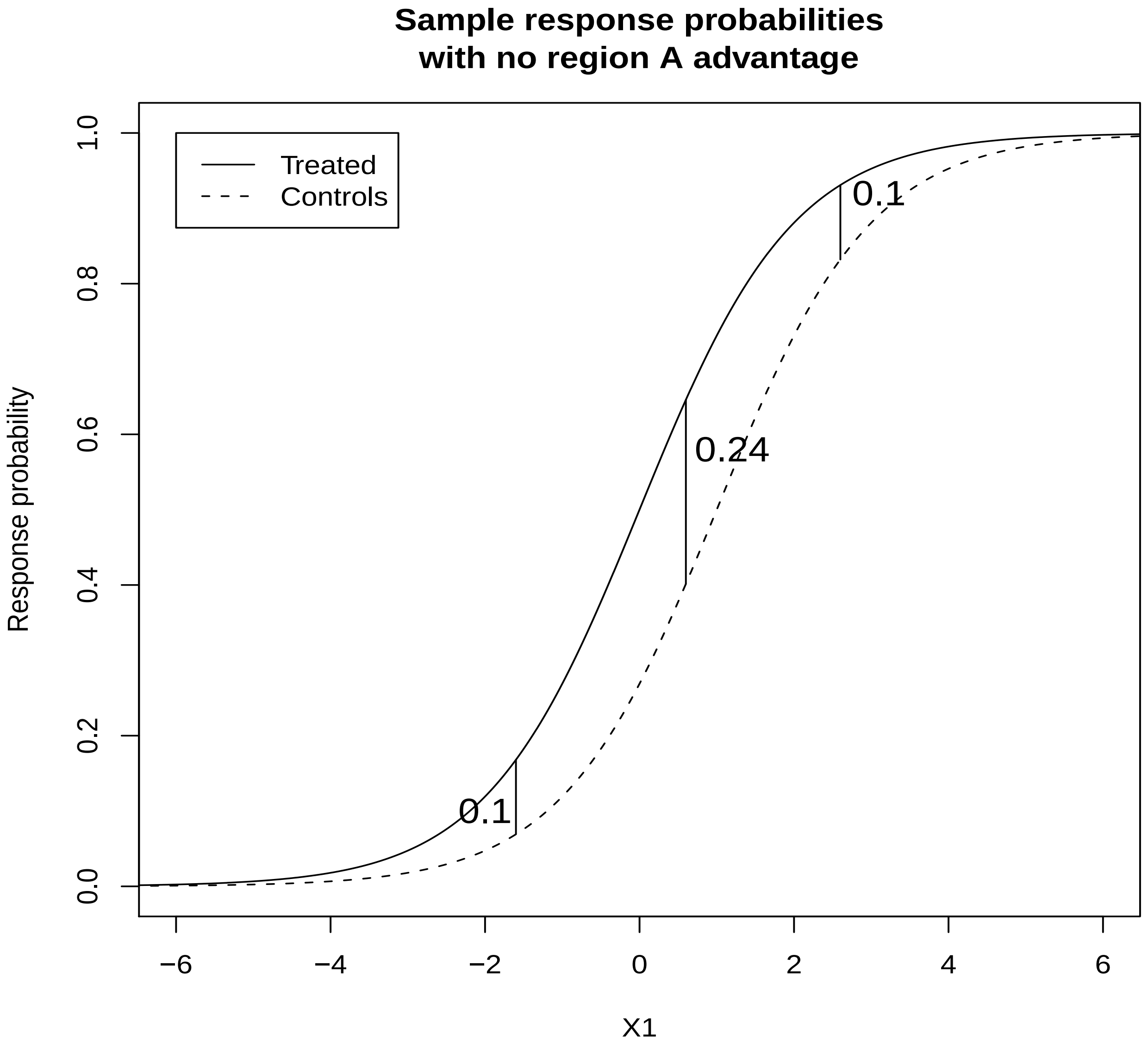

Response rate by X1
Treated
Controls
Response rate
0.0
0.2
0.4
0.6
0.8
1.0
(−Inf,−1.5]
(−1.5,−0.5]
(−0.5,0.5]
(0.5,1.5]
(1.5, Inf]
X1


Response rate by X2
Response rate
0.0
0.2
0.4
0.6
0.8
1.0
× Treated
● Controls
(−Inf,−1.5]
(−1.5,−0.5]
(−0.5,0.5]
(0.5,1.5]
(1.5, Inf]
X2

Enhanced treatment effect: no unique definition

- Difference in absolute risk, $P(Y = 1|T = 1, X) - P(Y = 1|T = 0, X)$
- Relative risk, $P(Y = 1|T = 1, X)/P(Y = 1|T = 0, X)$
- Difference in log-odds, $logit(P(Y = 1|T = 1, X)) - logit(P(Y = 1|T = 0, X))$

- Simple model
  - Treatment effect
  - X_1 is prognostic
  - No interaction
  - $logit(P(Y = 1|T, X_1)) = -1 + T + X_1$


Sample response probabilities
with no region A advantage
Treated
Controls
0.1
0.24
0.1
Response probability
0.0
0.2
0.4
0.6
0.8
1.0
X1
−6
−4
−2
0
2
4
6

Interactions in statistical models

- Main Effects Model
  - $logit(P(Y = 1|T, X)) = \alpha + \beta T + \gamma_1 X_1 + \gamma_2 X_2$
  - Note, no interaction on logit scale may have interaction on absolute risk scale
- Main Effects + Interaction
  - $logit(P(Y = 1|T, X)) = \alpha + \beta T + \gamma_1 X_1 + \gamma_2 X_2 + \delta T X_1$
  - $logit(P(Y = 1|T, X)) =$ $\alpha + \beta T + \gamma_1 X_1 + \gamma_2 X_2 + \delta T I(X \in A)$
- Need large sample sizes to find interactions

Naive method: Forward stepwise logistic regression

- Include
  - Main effects for T and X's
  - Interactions $X_j * X_k$
  - Interactions $T * X_j$
  - Interactions $T * X_j * X_k$
- Estimate $\hat{P}_{1i} = P(Y_i = 1|T_i = 1, X_i)$ and $\hat{P}_{0i} = P(Y_i = 1|T_i = 0, X_i)$ for each person i.
- New variable $Z_i = \hat{P}_{1i} - \hat{P}_{0i}$ is then created,
- Subjects i in group A if $Z_i > c$ (c=cutoff, say 0.15)

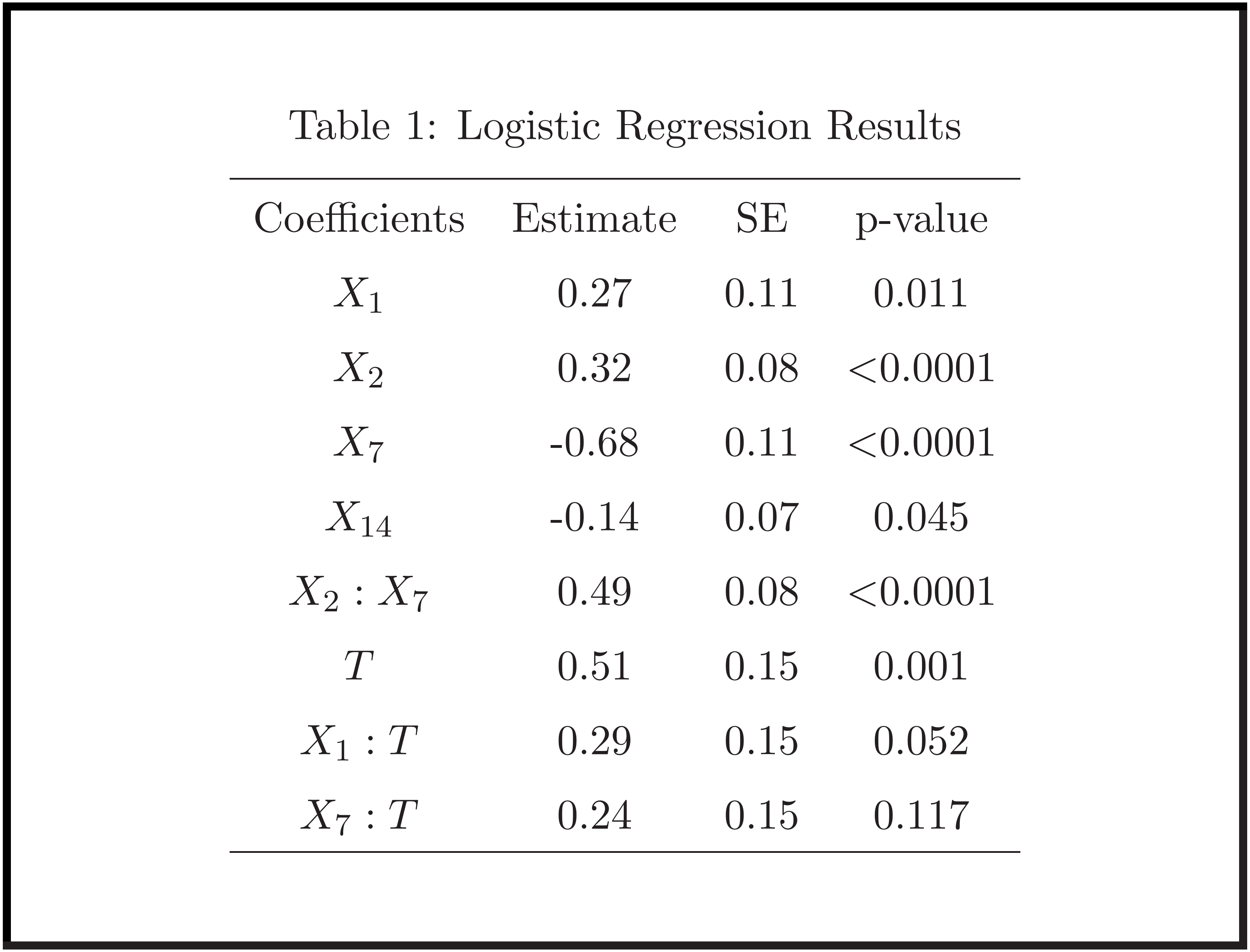


Table 1: Logistic Regression Results

Coefficients	Estimate	SE	p-value
X_1	0.27	0.11	0.011
X_2	0.32	0.08	<0.0001
X_7	-0.68	0.11	<0.0001
X_{14}	-0.14	0.07	0.045
$X_2 : X_7$	0.49	0.08	<0.0001
T	0.51	0.15	0.001
$X_1 : T$	0.29	0.15	0.052
$X_7 : T$	0.24	0.15	0.117

- X-by-T interaction found
- Region $\hat{A}$ estimated as $\hat{Z}_i > 0.168$

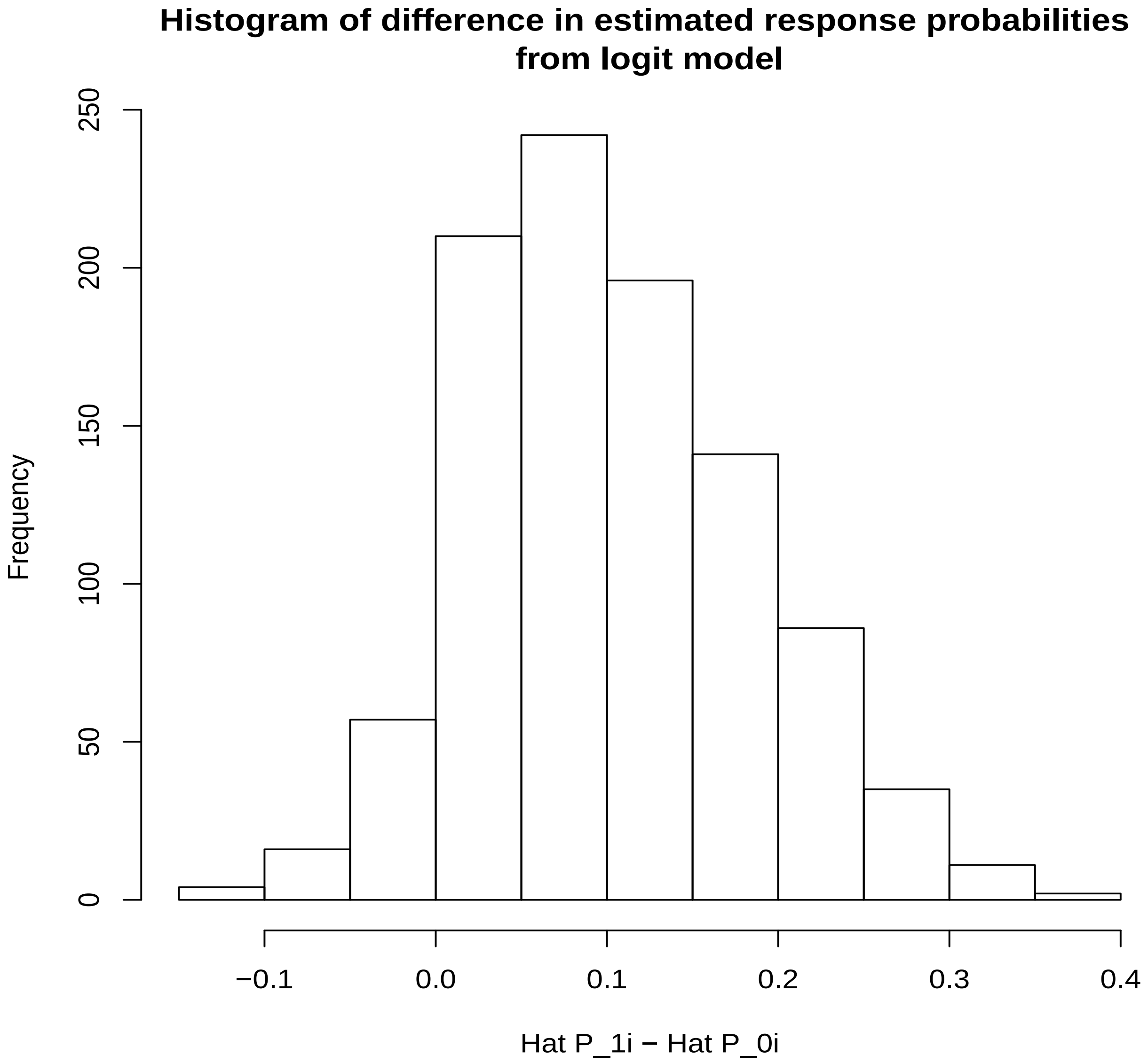

Histogram of difference in estimated response probabilities
from logit model
Frequency
0
50
100
150
200
250
−0.1
0.0
0.1
0.2
0.3
0.4
Hat P_1i − Hat P_0i

Table 2: Number of subjects in 4 cells (Logit)

	$\hat{A}$	not $\hat{A}$	
Treatment	110	390	500
Control	99	401	500
Overall	209	791	1000

Table 3: Response rate in 4 cells (Logit)

	Treatment	Control
$\hat{A}$	0.518	0.333
not $\hat{A}$	0.377	0.279
Overall	0.408	0.290

Table 4: How close is $\hat{A}$ to A (Logit)

	$\hat{A}$	not $\hat{A}$	
A	57	177	234
not A	152	614	766
Overall	209	791	1000

- Sensitivity = 0.24
- Specificity = 0.80
- Positive Predictive Value = 0.27
- Negative Predictive Value = 0.78

Virtual Twins method

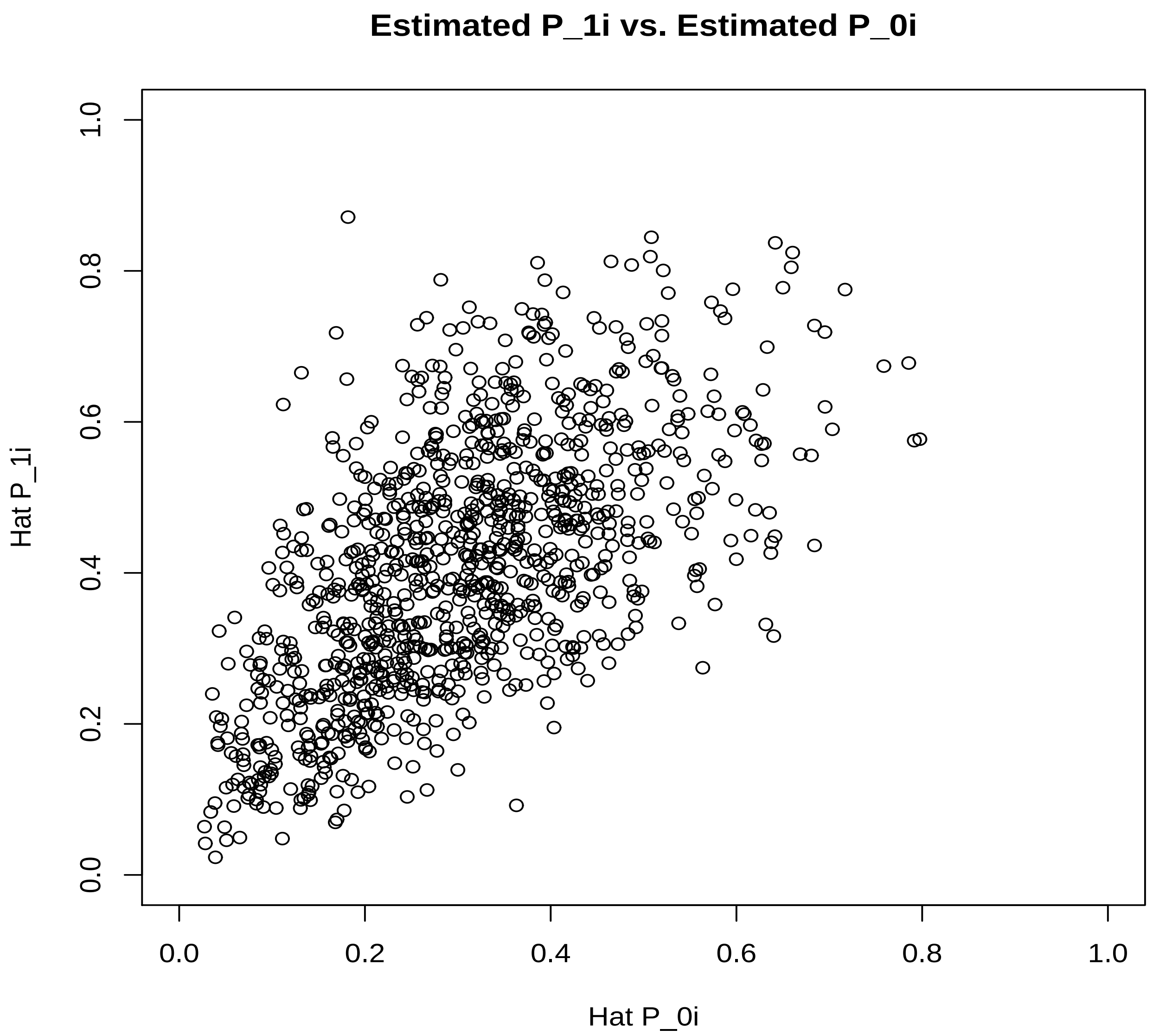
- For each person think about outcome if they got treatment and outcome if they got placebo
- Two steps
  - Step 1. Use Random Forests (RF) on all the data
  - Step 2. Run output from RF down a regression tree to find region A

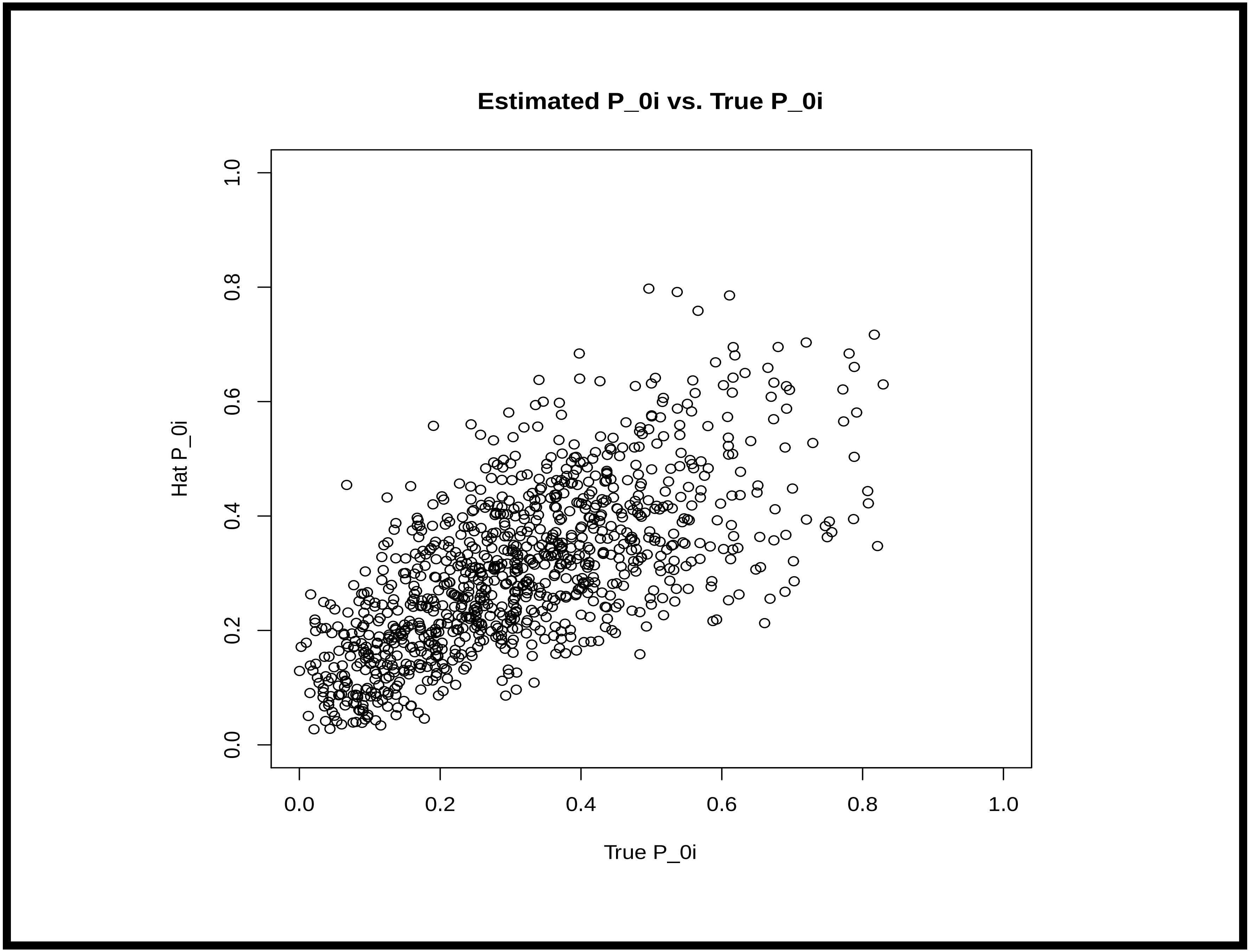
Random Forests: A type of non-parametric regression

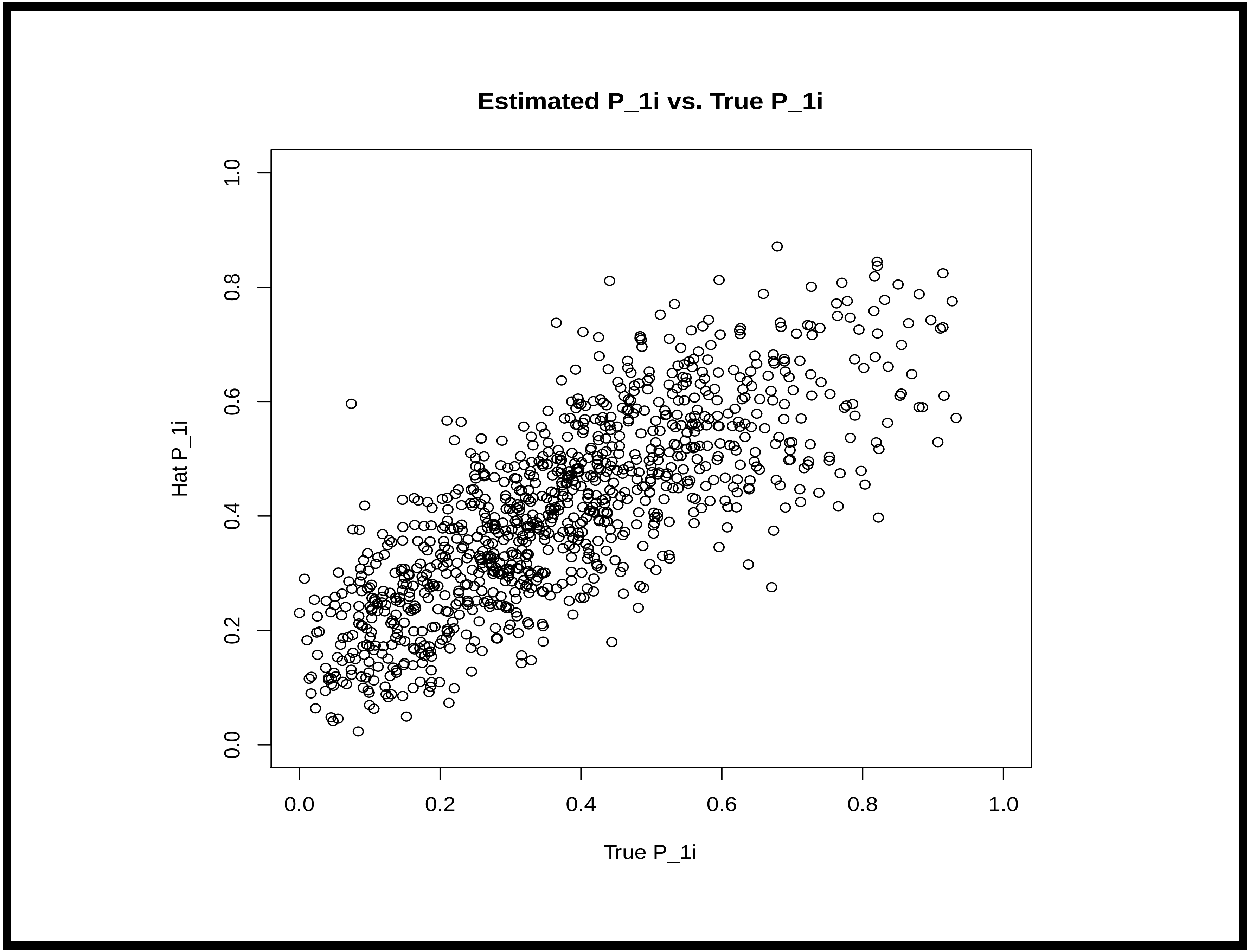
- A statistical learning algorithm for estimating function f(.) in the model $P(Y = 1) = f(T, X_1, .., X_p)$
- An ensemble method combining 250 trees
  - Combine many simple trees
  - Uses Bootstrap samples
  - Uses randon subsets of covariates at each split
  - Combine 250 predictions
- A black box
  - Input, values of T and X's
  - Output, estimate of $P(Y = 1|T, X)$

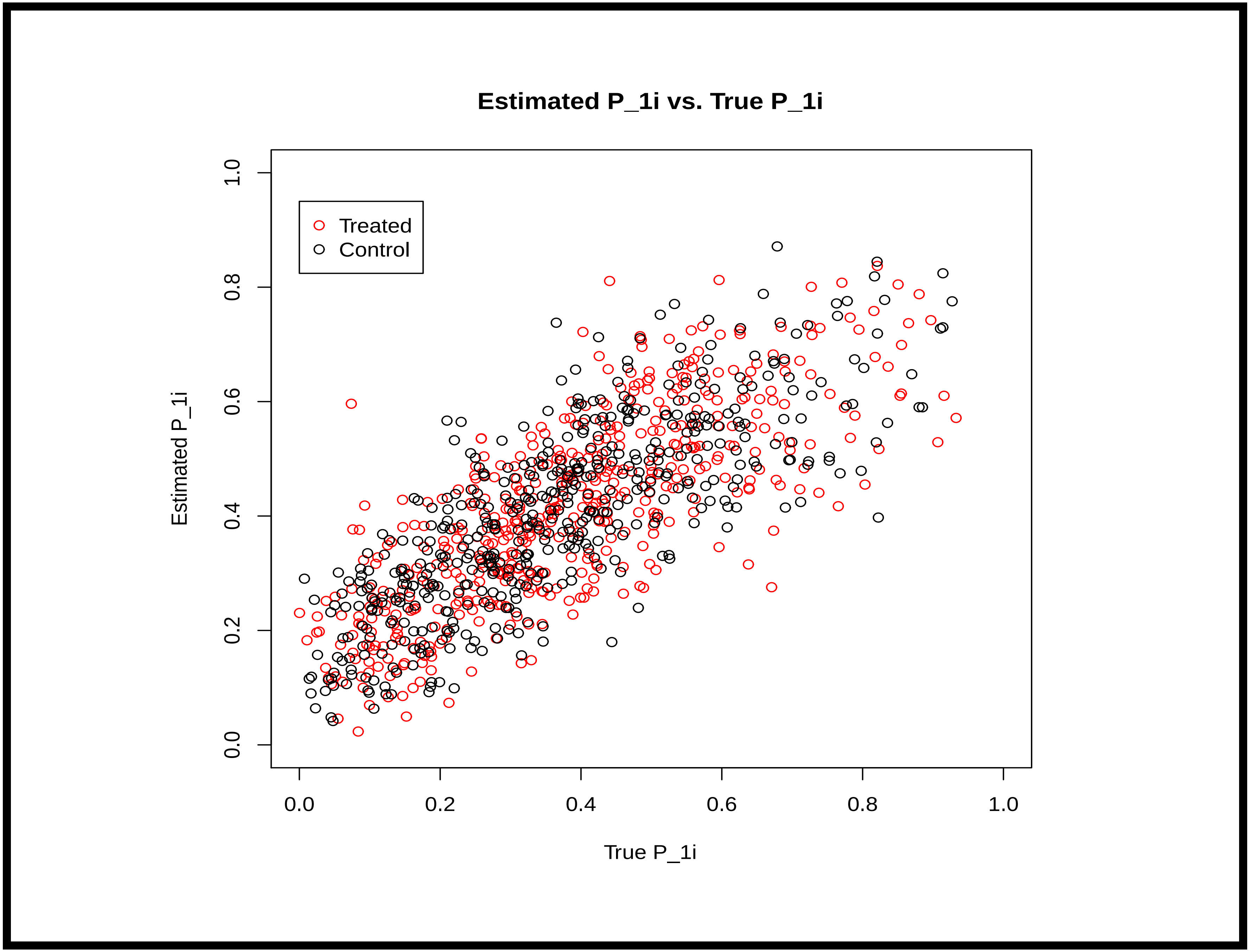
Step 1.

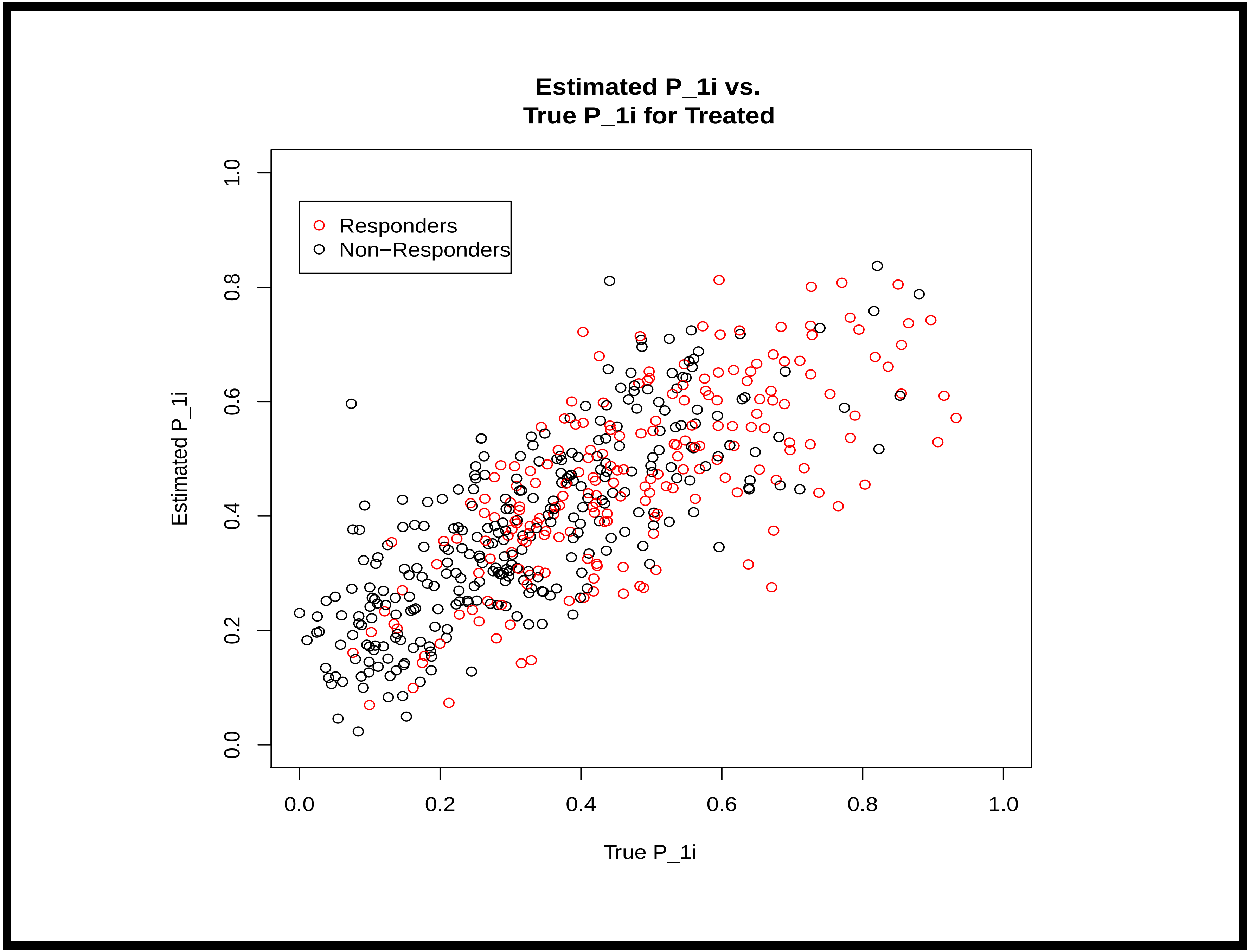
- Apply Random Forests to all the data
  - Covariates X, T, X*I(T=0), X*I(T=1)
  - Produces black box predictor
- For each subject apply predictor twice
  - Once to $(T = 1, X_{1i}, ..., X_{pi})$
  - Once to $(T = 0, X_{1i}, ..., X_{pi})$
  - Gives $\hat{P}_{1i} = P(Y_i = 1|T_i = 1, X_i)$ and $\hat{P}_{0i} = P(Y_i = 1|T_i = 0, X_i)$
- Form $Z_i = \hat{P}_{1i} - \hat{P}_{0i}$
- A measure of the treatment effect for subject i

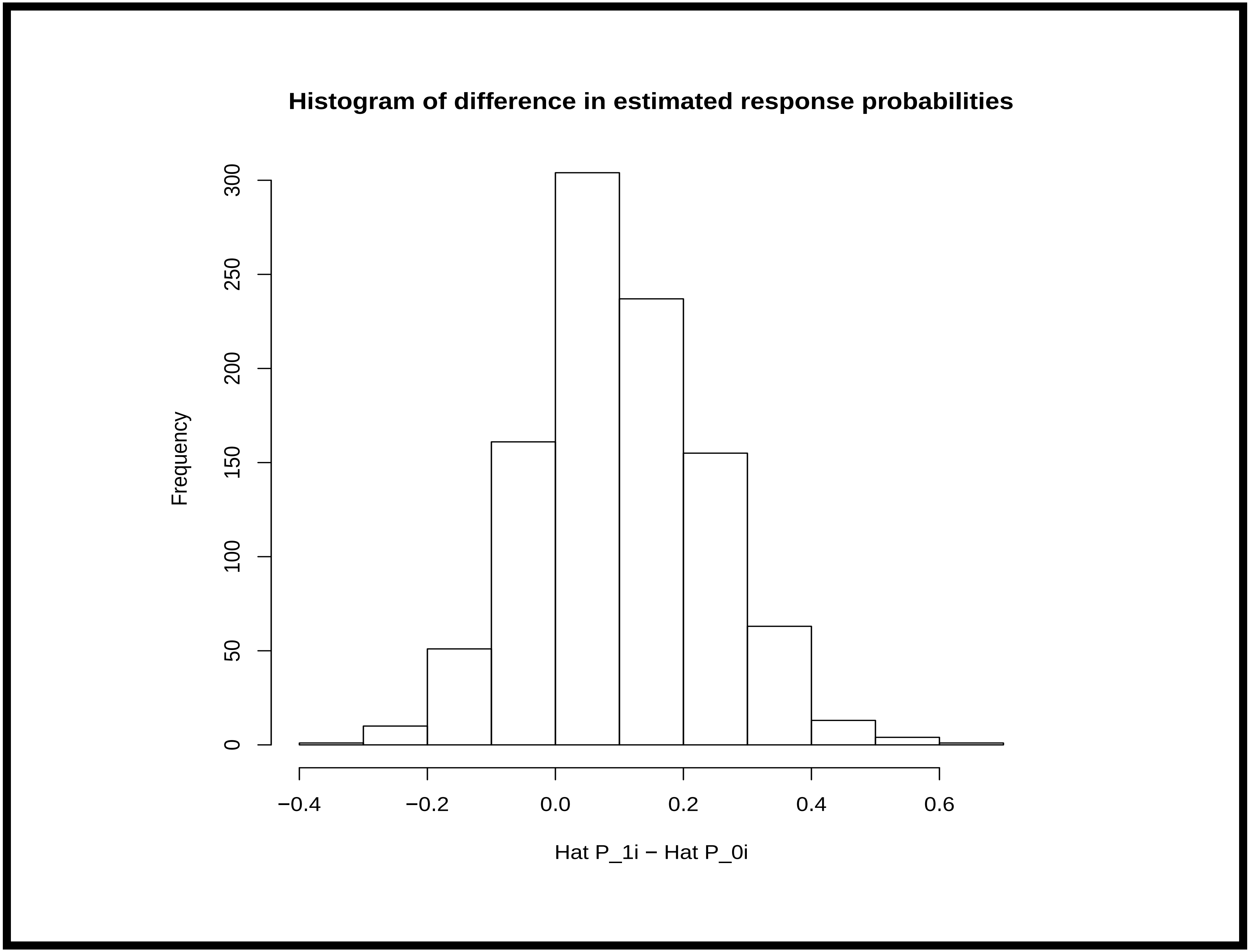

Estimated P_1i vs. Estimated P_0i
Hat P_1i
Hat P_0i
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0


Estimated P_0i vs. True P_0i
Hat P_0i
True P_0i
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0


Estimated P_1i vs. True P_1i
Hat P_1i
True P_1i
0.0
0.2
0.4
0.6
0.8
1.0
0.0
0.2
0.4
0.6
0.8
1.0

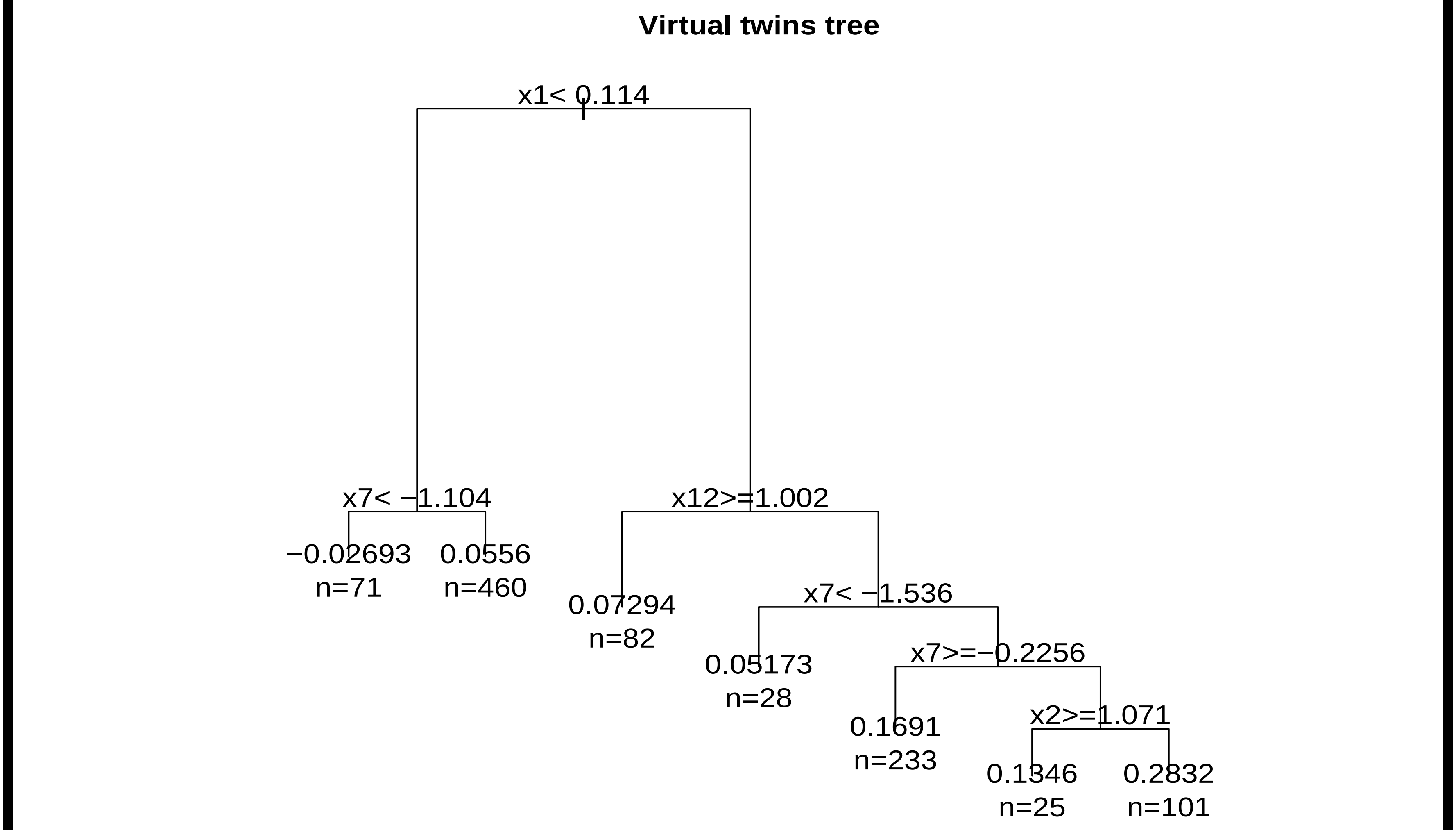

Estimated P_1i vs. True P_1i
○ Treated
○ Control
Estimated P_1i
0.0
0.2
0.4
0.6
0.8
1.0
True P_1i
0.0
0.2
0.4
0.6
0.8
1.0


Estimated P_1i vs.
True P_1i for Treated
Responders
Non−Responders
Estimated P_1i
0.0
0.2
0.4
0.6
0.8
1.0
True P_1i
0.0
0.2
0.4
0.6
0.8
1.0


Histogram of difference in estimated response probabilities
Frequency
0
50
100
150
200
250
300
−0.4
−0.2
0.0
0.2
0.4
0.6
Hat P_1i − Hat P_0i

Step 2. Regression Trees

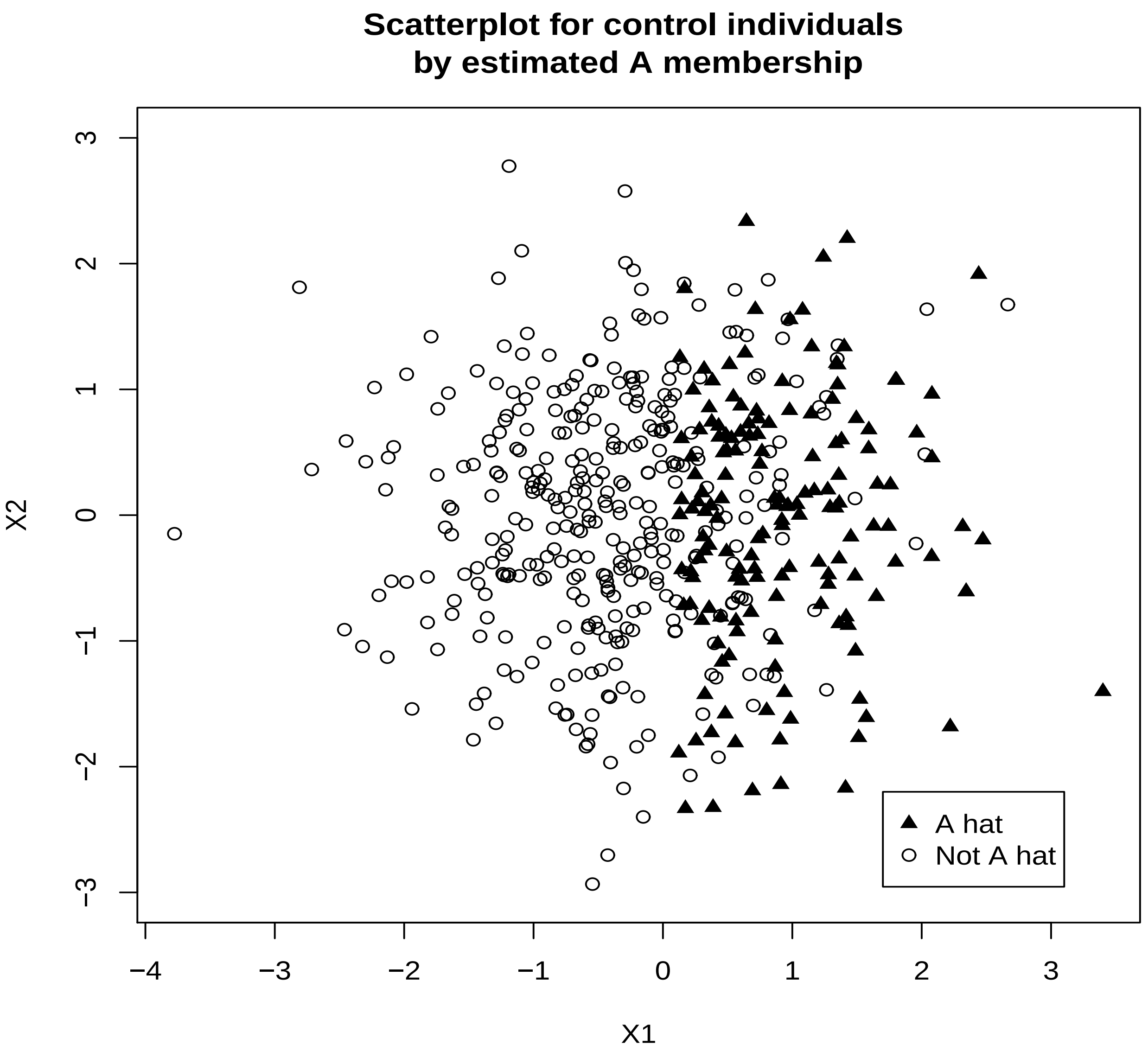
- Find small number of variables that are most associated with Z_i
- Estimate regression tree for Z_i with covariates $X_{1i}, ..., X_{pi}$
- The result, a small number of X's with cutpoints

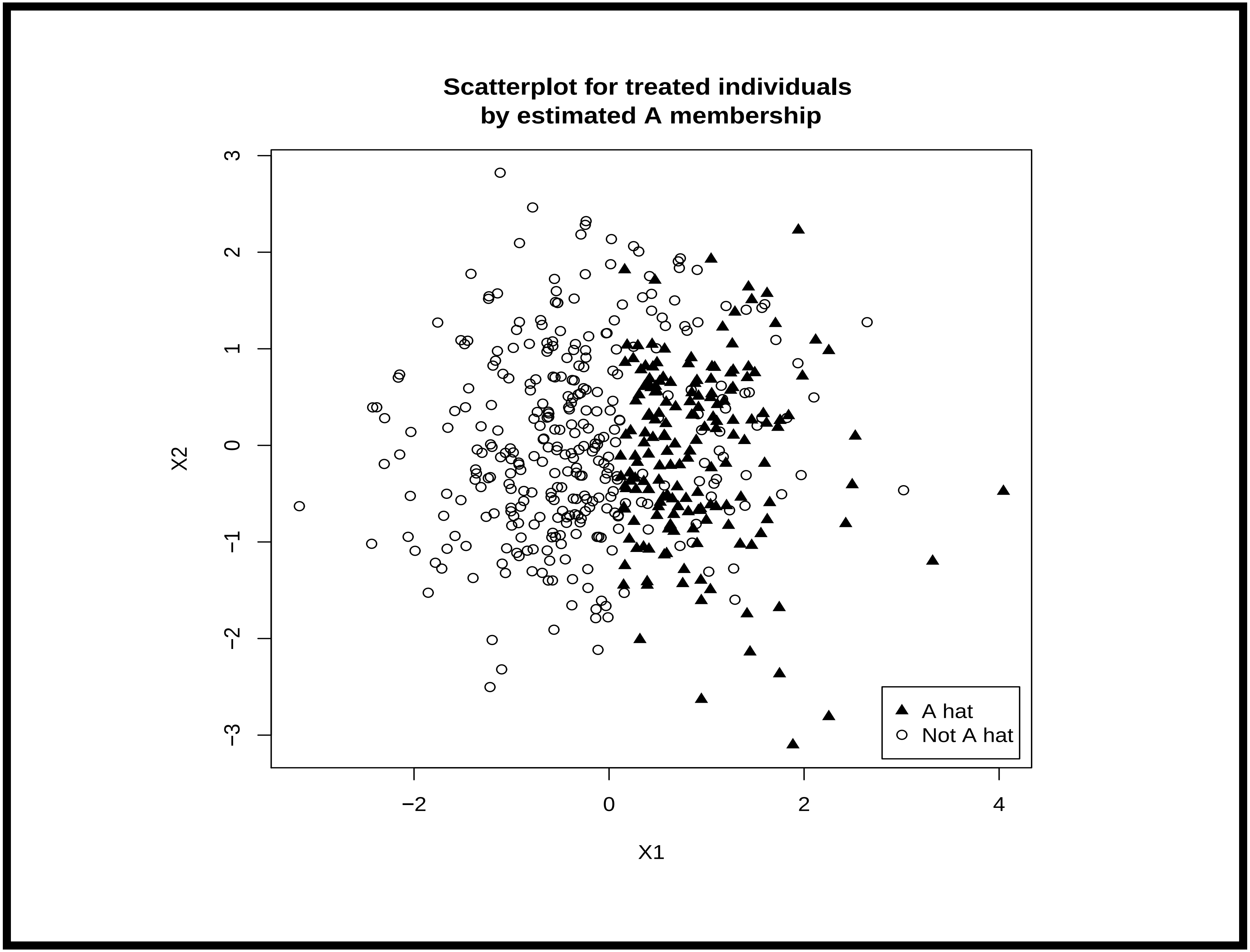

Virtual twins tree
x1< 0.114
x7< −1.104
x12>=1.002
−0.02693
n=71
0.0556
n=460
0.07294
n=82
x7< −1.536
0.05173
n=28
x7>=−0.2256
0.1691
n=233
x2>=1.071
0.1346
n=25
0.2832
n=101

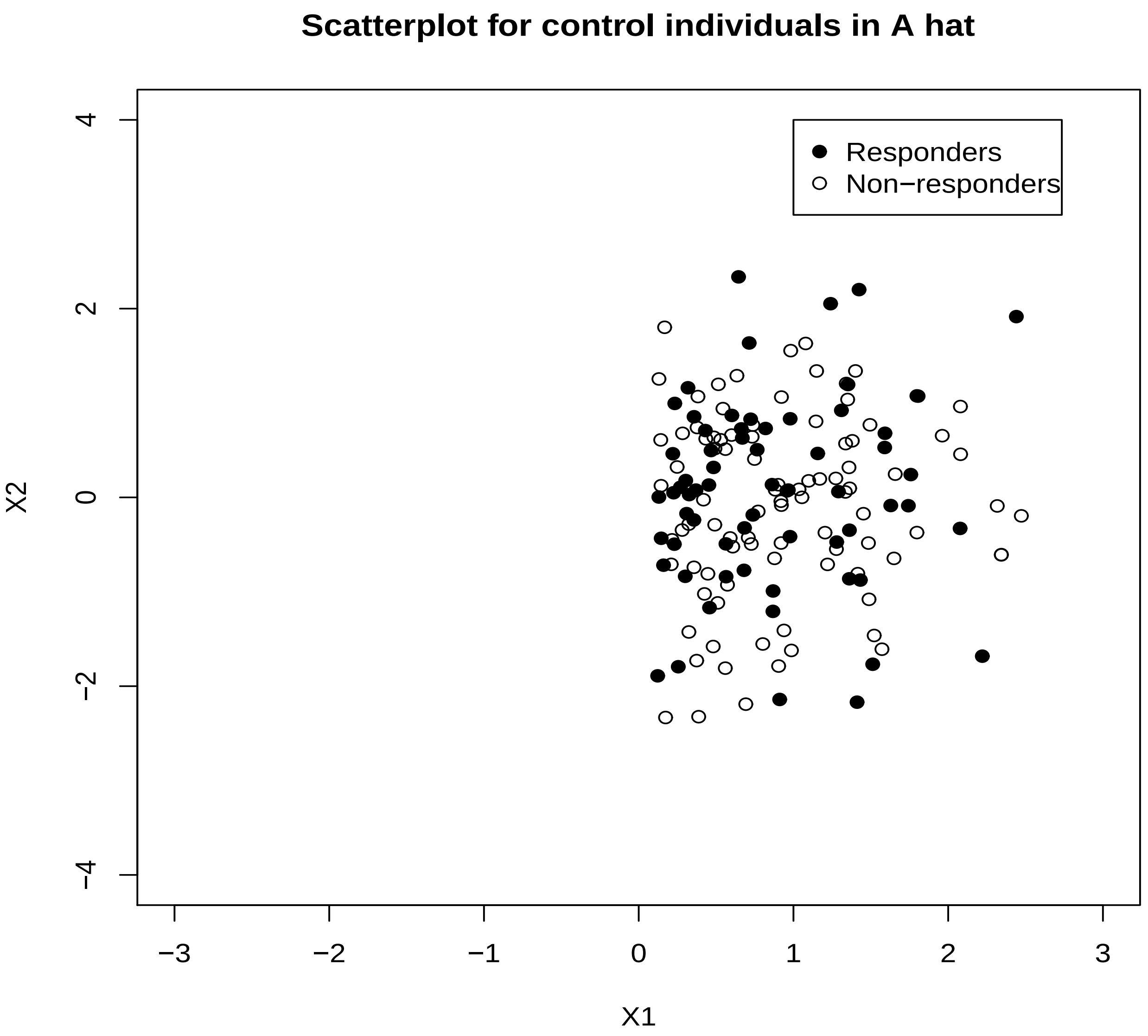
$\hat{A}$ classification

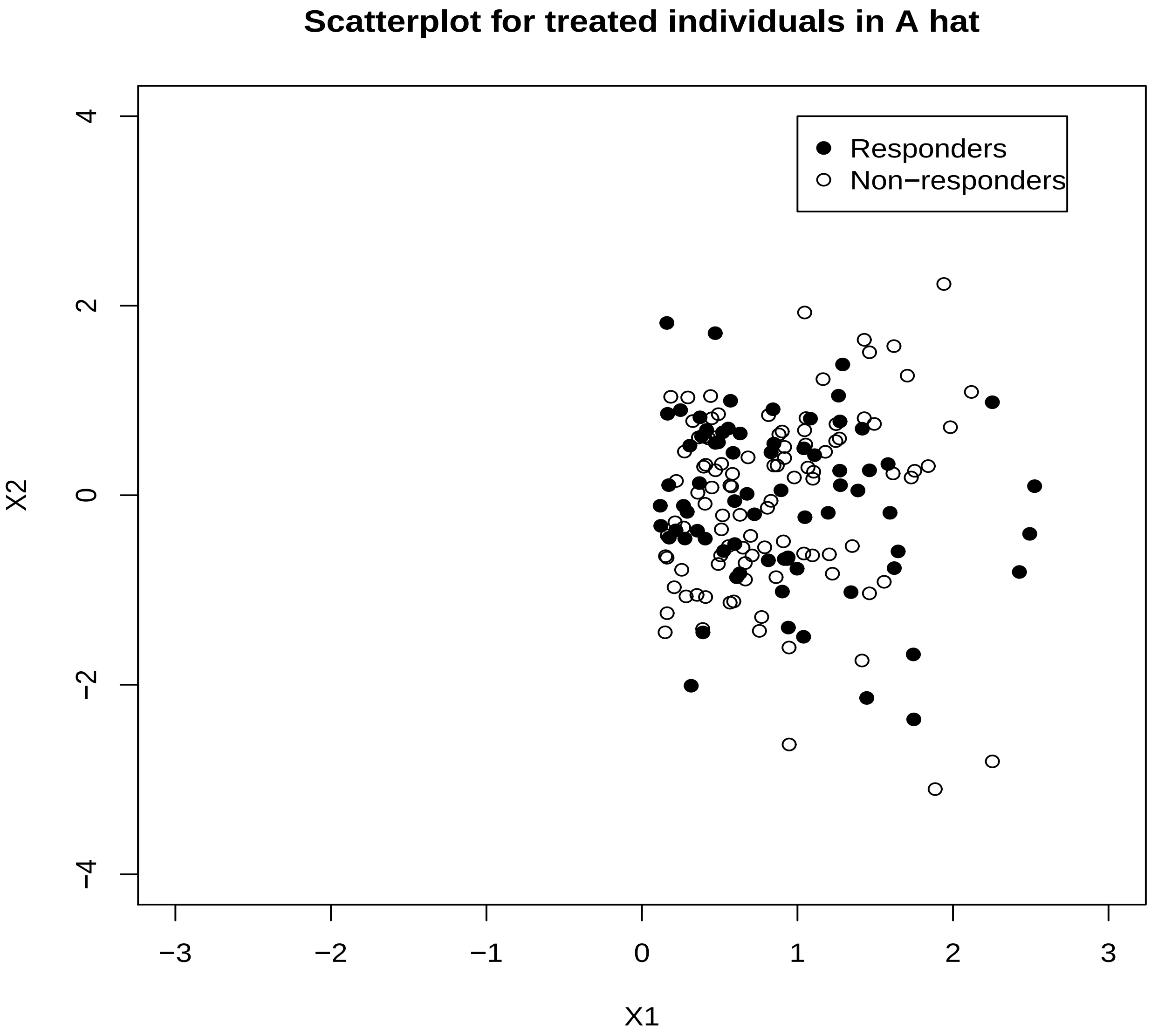
Table 5: Number of subjects in 4 cells (VT)

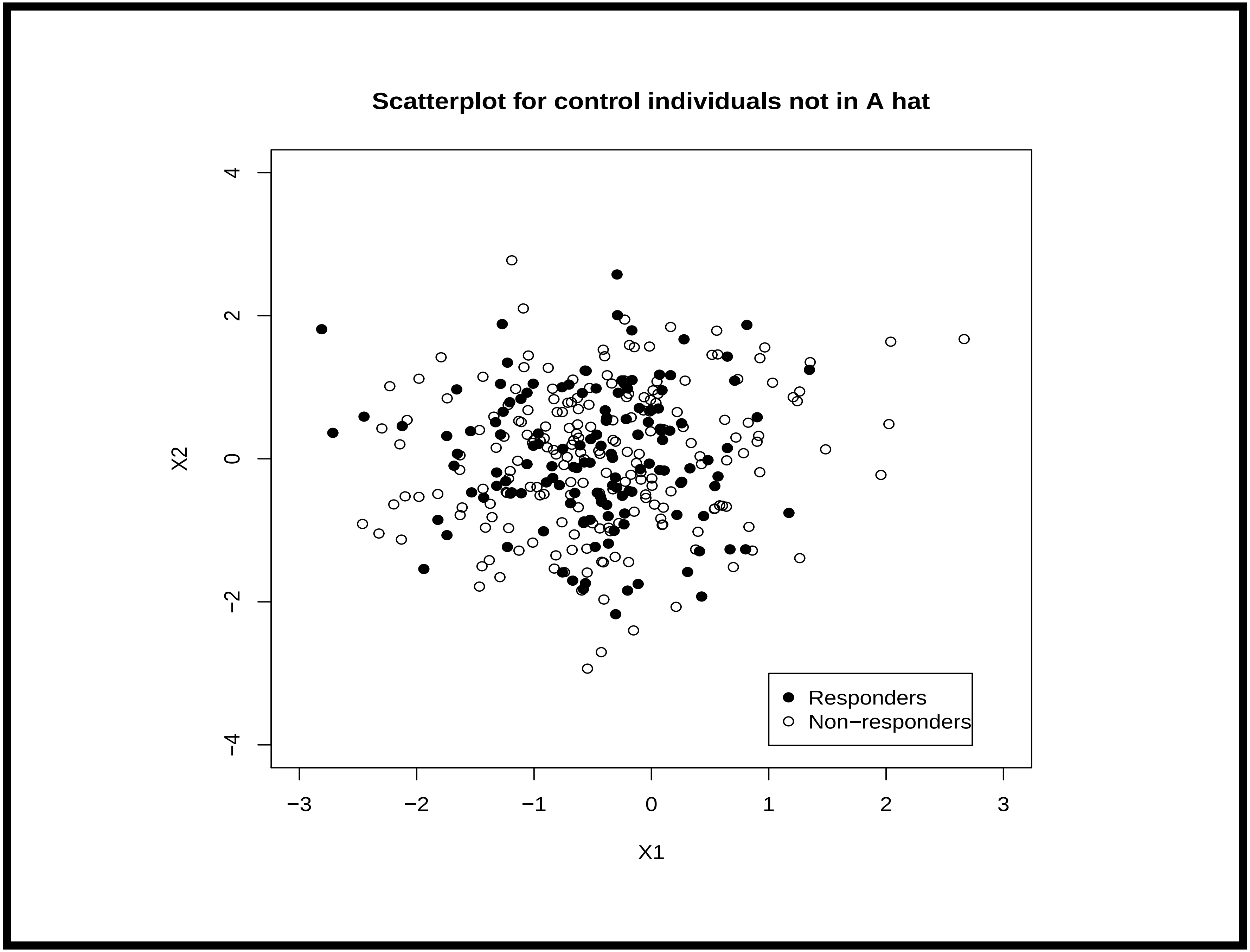
	$\hat{A}$	not $\hat{A}$	
Treatment	175	325	500
Control	159	341	500
Overall	334	666	1000


Scatterplot for control individuals
by estimated A membership
X2
X1
3
2
1
0
−1
−2
−3
−4
−3
−2
−1
0
1
2
3
▲ A hat
○ Not A hat


Scatterplot for treated individuals
by estimated A membership
X1
X2
−2
0
2
4
−3
−2
−1
0
1
2
3
▲ A hat
○ Not A hat


Scatterplot for control individuals in A hat
● Responders
○ Non−responders
X2
4
2
0
−2
−4
X1
−3
−2
−1
0
1
2
3


Scatterplot for treated individuals in A hat
Responders
Non−responders
X2
4
2
0
−2
−4
X1
−3
−2
−1
0
1
2
3


Scatterplot for control individuals not in A hat
X2
4
2
0
−2
−4
X1
−3
−2
−1
0
1
2
3
● Responders
○ Non−responders

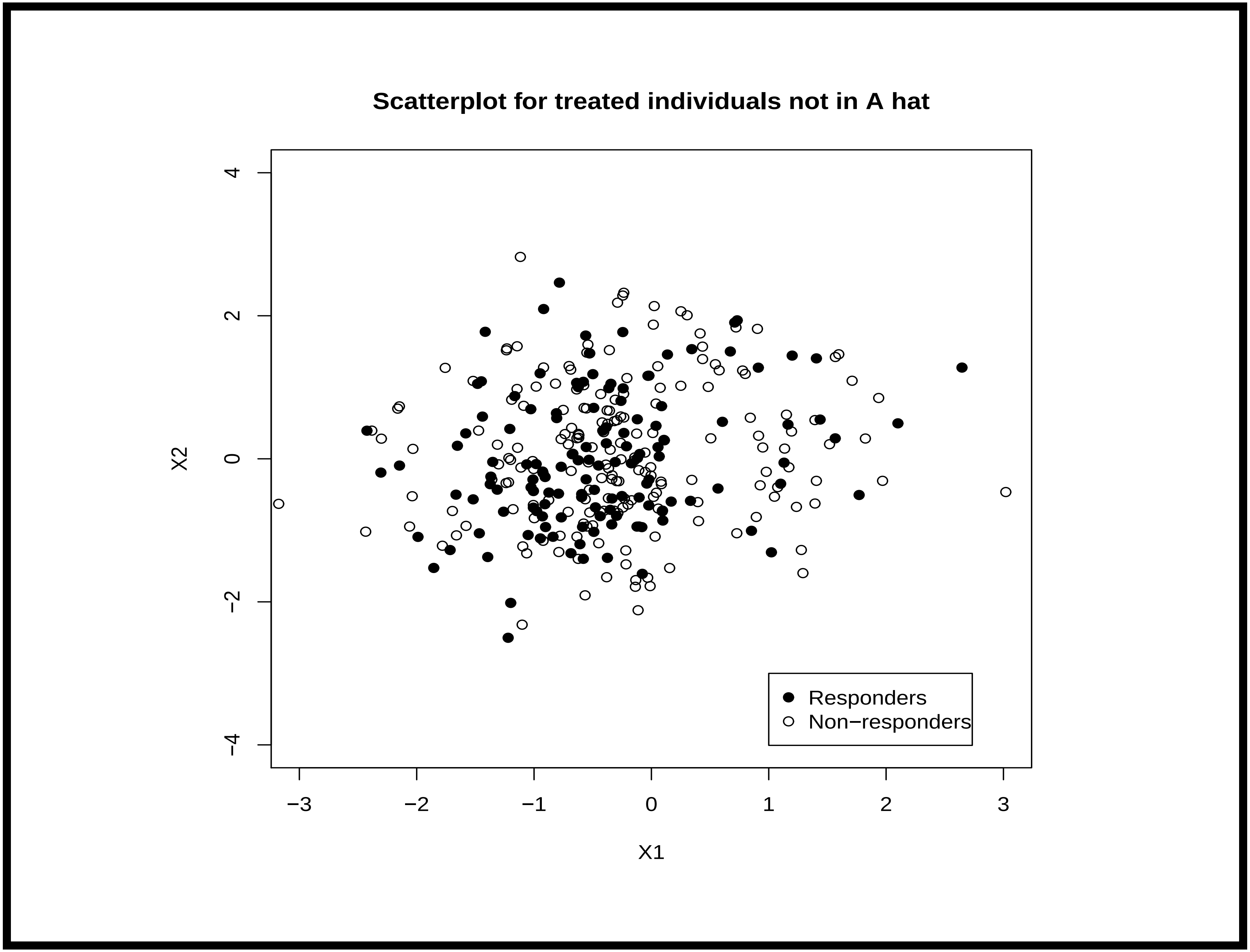

Scatterplot for treated individuals not in A hat
X1
X2
−3
−2
−1
0
1
2
3
−4
−2
0
2
4
● Responders
○ Non−responders

Table 6: Response rate in 4 cells (VT)

	Treatment	Control
$\hat{A}$	0.537	0.277
not $\hat{A}$	0.338	0.296
Overall	0.408	0.290

Properties of subgroup

- How big is $\hat{A}$?
- What X's does it depend on
- Quantify magnitude of enhanced treatment effect
- If know true A
  - Are we finding the correct X's
  - How close is $\hat{A}$ to true A
  - Sensitivity, Specificity
  - Positive Predictive Value, Negative Predictive Value

Table 7: How close is $\hat{A}$ to A (VT)

	$\hat{A}$	not $\hat{A}$	
A	159	75	234
not A	175	591	766
Overall	334	666	1000

- Sensitivity = 0.68
- Specificity = 0.77
- Positive Predictive Value = 0.48
- Negative Predictive Value = 0.89

Metrics for enhanced treatment effects

$$\begin{aligned} Q(\hat{A}) &= (P(Y=1|T=1, X \in \hat{A}) - P(Y=1|T=0, X \in \hat{A})) \\ &\quad -(P(Y=1|T=1) - P(Y=1|T=0)) \end{aligned}$$

Table 8: Response rate in 4 cells (VT)

	Treatment	Control
$\hat{A}$	0.537	0.277
not $\hat{A}$	0.338	0.296
Overall	0.408	0.290

- $\hat{Q}(\hat{A})_{VT}$ = (0.537-0.277)-(0.408-0.290)=0.142

Table 9: Response rate in 4 cells (Logit)

	Treatment	Control
$\hat{A}$	0.518	0.333
not $\hat{A}$	0.377	0.279
Overall	0.408	0.290

- $\hat{Q}(\hat{A})_{Logit}$ = (0.518-0.333) - (0.408-0.290) = 0.067

- Notation
  - $Q(\hat{A})$ = true value of Q for $\hat{A}$
  - $\hat{Q}(\hat{A})$ = estimate of Q for $\hat{A}$
  - Want estimates to have low bias and small variability (small SE)

Resubstitution estimates

- $\hat{Q}(\hat{A})_{VT} = 0.142$
- $\hat{Q}(\hat{A})_{Logit} = 0.067$
- Almost certainly optimistically biased estimates
- Need honest estimate of $Q(\hat{A})$
  - What would $Q(\hat{A})$ be with this $\hat{A}$ in the next very large trial

Methods of estimating $Q(\hat{A})$

- Resubstitution method
- Simulate new data
- Cross-validation of $\hat{P}_{1i}$ and $\hat{P}_{0i}$.
- Full Cross-validation

Simulate new data

- Simulate new data, that ”looks like” original data, but is ”independent”
- Generate binary Y_i using either $\hat{P}_{1i}$ or $\hat{P}_{0i}$.
- if $T_i = 1$ then $Y_i^* \sim Bernoulli(\hat{P}_{1i})$
- if $T_i = 0$ then $Y_i^* \sim Bernoulli(\hat{P}_{0i})$
- Calculate $\hat{Q}(\hat{A})$ from these new data
- Repeat many times and average
- $\hat{Q}(\hat{A})_{VT} = 0.095$, $\hat{Q}(\hat{A})_{Logit} = 0.103$

Cross-validation of $\hat{P}_{1i}$ and $\hat{P}_{0i}$.

- Same as Simulate New Data, except $\hat{P}_{1i}$ and $\hat{P}_{0i}$ are derived after cross-validation
- Take 9/10 of data, run Random Forest (or Logit model with forward selection), predict for left out 1/10
- Repeat 10 times
- $\hat{Q}(\hat{A})_{VT} = 0.124$, $\hat{Q}(\hat{A})_{Logit} = 0.081$

Full Cross-validation

- Take 9/10 of data
- Find region $\hat{A}_k$
- Find $\hat{Q}(\hat{A}_k)$ for left out 1/10
- Repeat 10 times
- Combine 10 separate $\hat{Q}(\hat{A}_k)$ to final $\hat{Q}(\hat{A})$
- $\hat{Q}(\hat{A})_{VT} = 0.089$, $\hat{Q}(\hat{A})_{Logit} = -0.071$

Table 10: Summary of estimates of $Q(\hat{A})$

Estimation Method	Virtual Twins	Logit
Resubstitution	0.142	0.067
Simulate new data	0.095	0.103
Cross-validation of $\hat{P}_{1i}$ and $\hat{P}_{0i}$	0.124	0.081
Full Cross-validation	0.089	-0.071
True value of $Q(\hat{A})$	0.031	-0.008
True value of $Q(A)$	0.133	0.133

Sampling variability of $\hat{Q}(\hat{A})$

- Also desirable to attach standard errors to $\hat{Q}(\hat{A})$
- We propose the following:
  1. Simulate many datasets using $\hat{P}_{1i}$ and $\hat{P}_{0i}$ from random forest/logistic method
  2. For each data set, estimate a new $\hat{A}$ and calculate $\hat{Q}(\hat{A})$
  3. Estimated standard error equals the standard deviation of these $\hat{Q}(\hat{A})$

Table 11: Standard errors for $\hat{Q}(\hat{A})$

	Virtual Twins		Logit	
Method	Est.	SE	Est.	SE
Resubstitution	0.142	0.094	0.067	0.072
Sim. new data	0.095	0.055	0.103	0.077

Null distribution of $\hat{Q}(\hat{A})$ and p-values

- Null distribution $\equiv$ no region of enhanced treatment effect

- Null should allow possibility of main effects for X and T, but with no interaction

- We propose the following:
  1. Define $V_i = logit(\hat{P}_{1i}) - logit(\hat{P}_{0i})$ and $\overline{V} = \frac{1}{n}\sum V_i$
  2. Define $\hat{P}^N_{1i} = expit(\frac{logit(\hat{P}_{1i})+logit(\hat{P}_{0i})}{2} + \frac{\overline{V}}{2})$, and $\hat{P}^N_{0i} = expit(\frac{logit(\hat{P}_{1i})+logit(\hat{P}_{0i})}{2} - \frac{\overline{V}}{2})$
  3. Simulate many datasets using $\hat{P}^N_{1i}$ and $\hat{P}^N_{0i}$

4. For each data set, first estimate $\hat{A}$ and calculate $\hat{Q}(\hat{A})$
5. P-value is the fraction of these $\hat{Q}(\hat{A})$ that are larger than the observed $\hat{Q}(\hat{A})$
6. If original $\hat{A}$ is empty, take p-value to be 0.5.

Table 12: P-values for $\hat{Q}(\hat{A})$

Method	Virtual Twins	Logit
Resubstitution	0.335	0.085
Simulate new data	0.295	0.110

Simulation study

- Generate multiple datasets from $logit(P(Y = 1|T, X) = \alpha + \beta T + \gamma h(X) + \theta T I(X \in A)$
- A is a known region in the design space defined by a small number of X's
- Possible factors to consider
  - sample size, number of X's, correlation between X's,
  - size of true A, number of X's that determine true A, strength of enhanced treatment effect

Properties of subgroup

- Simulations (know true A)
  - Are we finding the correct X's
  - How big is $\hat{A}$?
  - How close is $\hat{A}$ to true A
  - Sensitivity, Specificity
  - Positive Predictive Value, Negative Predictive Value
  - Accuracy of estimates of $Q(\hat{A})$

- Used model:
  $logit(P(Y = 1)) =$
  $-1 + 0.5X_1 + 0.5X_2 - 0.5X_7 + 0.1T + 0.5X_2X_7 + \theta T I\{X \in A\}$
- 100 datasets generated
- For each, $n = 1000$

Table 13: Selected X's, $\theta = 0.75$, $A = \{X_1 > 0, X_2 < 0\}$

	Logit	Virtual Twins
Mean (# Unique X's)	0.98	3.45
SD (# Unique X's)	0.99	0.91
Pct. Found X_1 (int)	27	87
Pct. Found X_2 (int)	31	72
Pct. Found X_7 (main)	15	59
Pct. Found X_3 (null)	0	13
Pct. Found X_1 at top of tree		71
Pct. Found X_1 top 2 of tree		62
Pct. Found no int/tree	39	0

Table 14: Compare A to $\hat{A}$, $\theta = 0.75$, $A = \{X_1 > 0, X_2 < 0\}$

	Logit	Virtual Twins
percent $\hat{A}$ empty	39	7
size of $\hat{A}$ (median):	189	204
Sensitivity	0.29	0.44
Specificity	0.89	0.87
PPV	0.29	0.50
NPV	0.80	0.83
AUC	0.55	0.75

Table 15: Q Estimates, $\theta = 0.75$, $A = \{X_1 > 0, X_2 < 0\}$

	Logit		Virtual Twins	
	Mean	SD	Mean	SD
$Q(A)$	0.11		0.11	
$Q(\hat{A})$	0.027	0.03	0.050	0.042
$\hat{Q}(\hat{A})$:				
Resub	0.055	0.061	0.150	0.077
SimNewDat	0.061	0.052	0.104	0.048
Cr.Val	0.049	0.047	0.132	0.064
FullCr.Val	-0.016	0.107	0.110	0.088

Table 16: Selected X's, Null case

	Logit	Virtual Twins
Mean (# Unique X's)	0.18	3.34
SD (# Unique X's)	0.58	1.13
Pct. Found X_1 (int)	6	72
Pct. Found X_2 (int)	6	62
Pct. Found X_7 (main)	2	70
Pct. Found X_3 (null)	0	9
Pct. Found no tree/int	90	0

Table 17: Properties of $\hat{A}$, Null case

	Logit	Virtual Twins
percent $\hat{A}$ empty	89	16
Specificity	0.97	0.84

Table 18: Q Estimates, Null case

	Logit		Virtual Twins	
	Mean	SD	Mean	SD
$Q(A)$	0.003		0.003	
$Q(\hat{A})$	0.001	0.007	0.005	0.036
$\hat{Q}(\hat{A})$:				
Resub	0.011	0.033	0.128	0.096
SimNewDat	0.012	0.034	0.088	0.054
Cr.Val	0.008	0.024	0.112	0.075
FullCr.Val	-0.018	0.145	0.092	0.106

Lots of possibilities for adaptation and extension

- Replace Random Forest with other predictor
- Generalize to high dimensional data (genomic/genetic)
- Build in to the study design
- Vary thresholds to make smaller or larger $\hat{A}$
- Use different metrics for $Q(A)$
- Use different definitions of Z_i