What are the effects of "Bagging"? Some experimental and theoretical results

Andreas Buja
Professor, Statistics
The Wharton School
University of Pennsylvania

Werner Stuetzle *
Professor, Statistics
Adjunct Professor, CSE
University of Washington

* Supported by NSF grant DMS-9803226.
Research performed while on sabbatical at AT&T Labs – Research

Research motivated by Friedman & Hall paper ``On Bagging and Nonlinear Estimation'' (available on the Web)
and counter-example to one of F & H's claims due to Yoram Gatt.

The generic prediction problem

Given: *Training sample* $\mathcal{X} = \{(\underline{x}_1, y_1), \ldots, (\underline{x}_n, y_n)\}$
assumed to be iid obs of $(\underline{X}, Y)$, where
$\underline{X}$: vector of *predictor variables*
Y: *response variable*

Goal: Generate *prediction rule* (or *model*) $p(\underline{x}; \mathcal{X})$
to predict value of response Y
for predictor value $\underline{x}$

Classification and Regression Trees (Cart)

- Predict Y for predictor value $\underline{x}_0$ by average response of training observations in a neighborhood of $\underline{x}_0$.
- Neighborhoods are axis-parallel rectangles forming a partitioning of the predictor space $\Rightarrow$ model is piecewise constant over rectangles.
- Partitioning is constructed by a greedy search algorithm attempting to minimize the average squared prediction error for the training sample.

(Details not important here)

Bagging (Breiman 1996)

- Draw Bootstrap samples $\mathcal{X}_1, \ldots, \mathcal{X}_B$ from training sample
- Generate prediction rules $p(\underline{x}; \mathcal{X}_1), \ldots, p(\underline{x}; \mathcal{X}_B)$ from the Bootstrap samples
- Average the rules: $p^b(\underline{x}; \mathcal{X}) = \mathbf{ave}\ (p(\underline{x}, \mathcal{X}_1), \ldots, p(\underline{x}; \mathcal{X}_B))$

For euclidean response: $\mathbf{ave}$ = mean
For categorical response: $\mathbf{ave}$ = majority vote

Empirical evaluation:

Bagging effective in reducing the error rate of Cart classification and regression.

Illustration of Bagging

$X \sim U[0, 1]$
$Y = X + \epsilon$ with $\epsilon \sim N(0, 1)$
$n = 200$

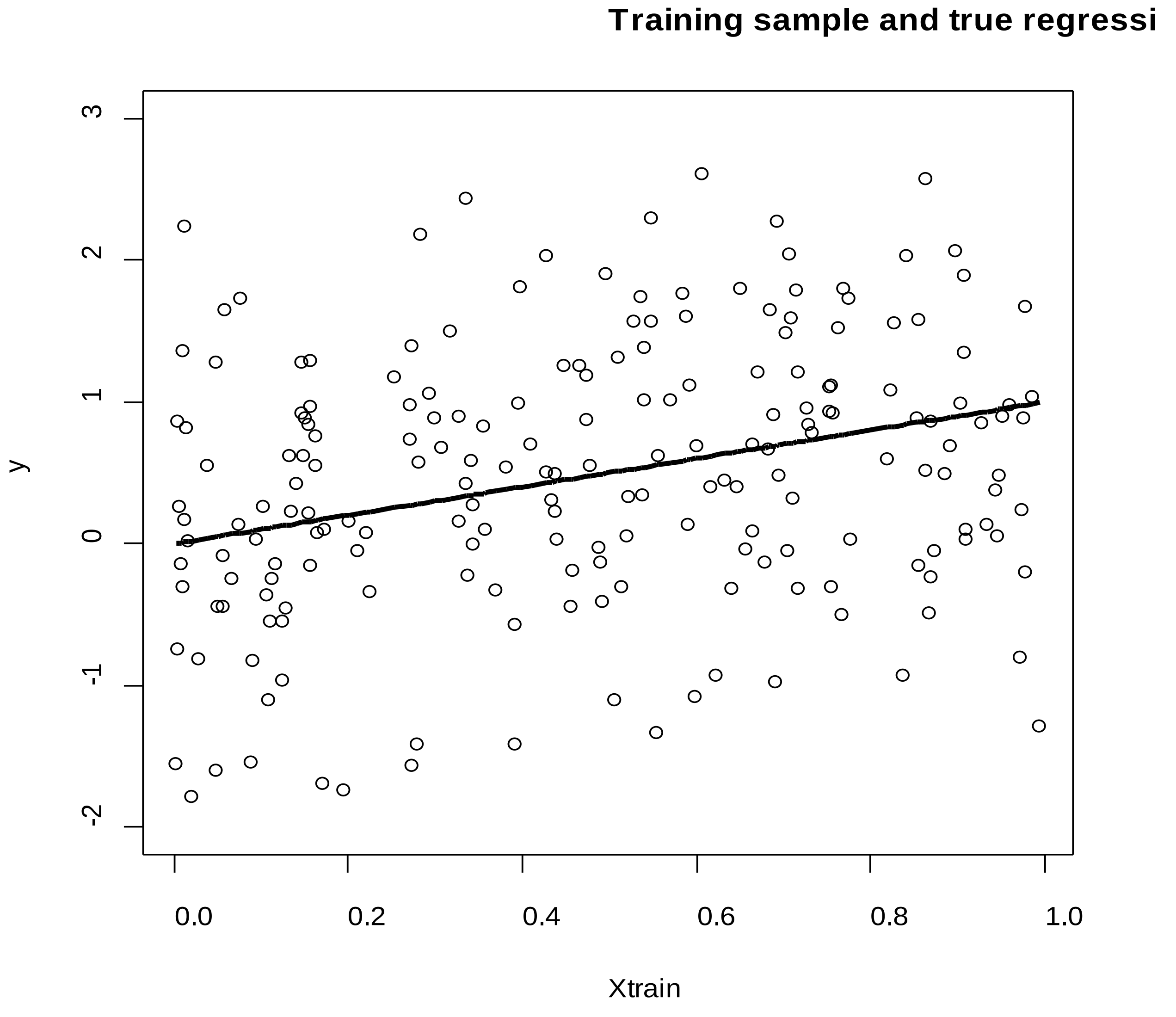
Partitition predictor space into two "rectangles."

Draw 50 resamples for bagging.

(Simple example, but illustrates all the effects of bagging)

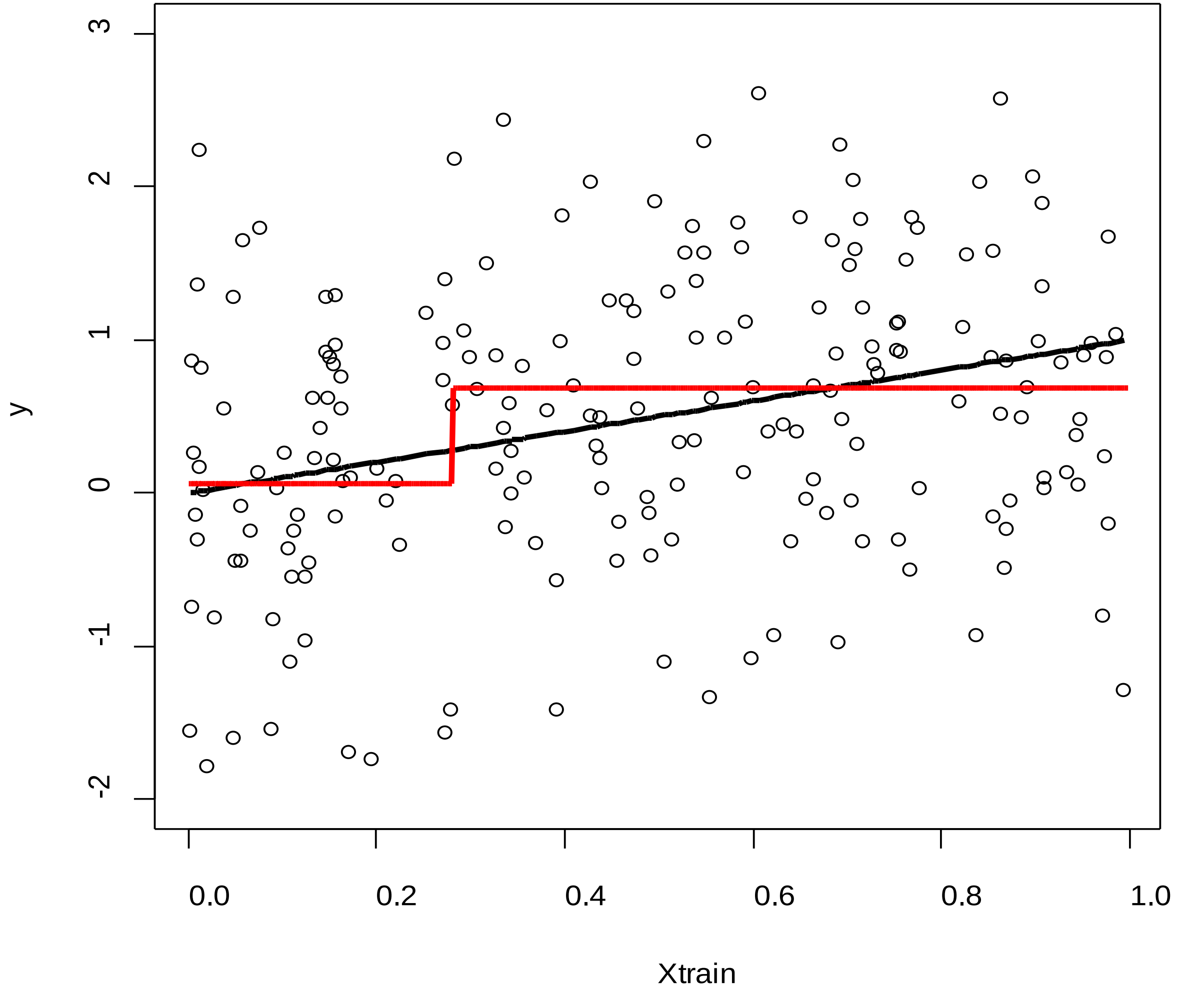
First consider a single training sample.

- Look at Cart model for training sample and for 10 resamples.
- Then compare bagged and unbagged models.

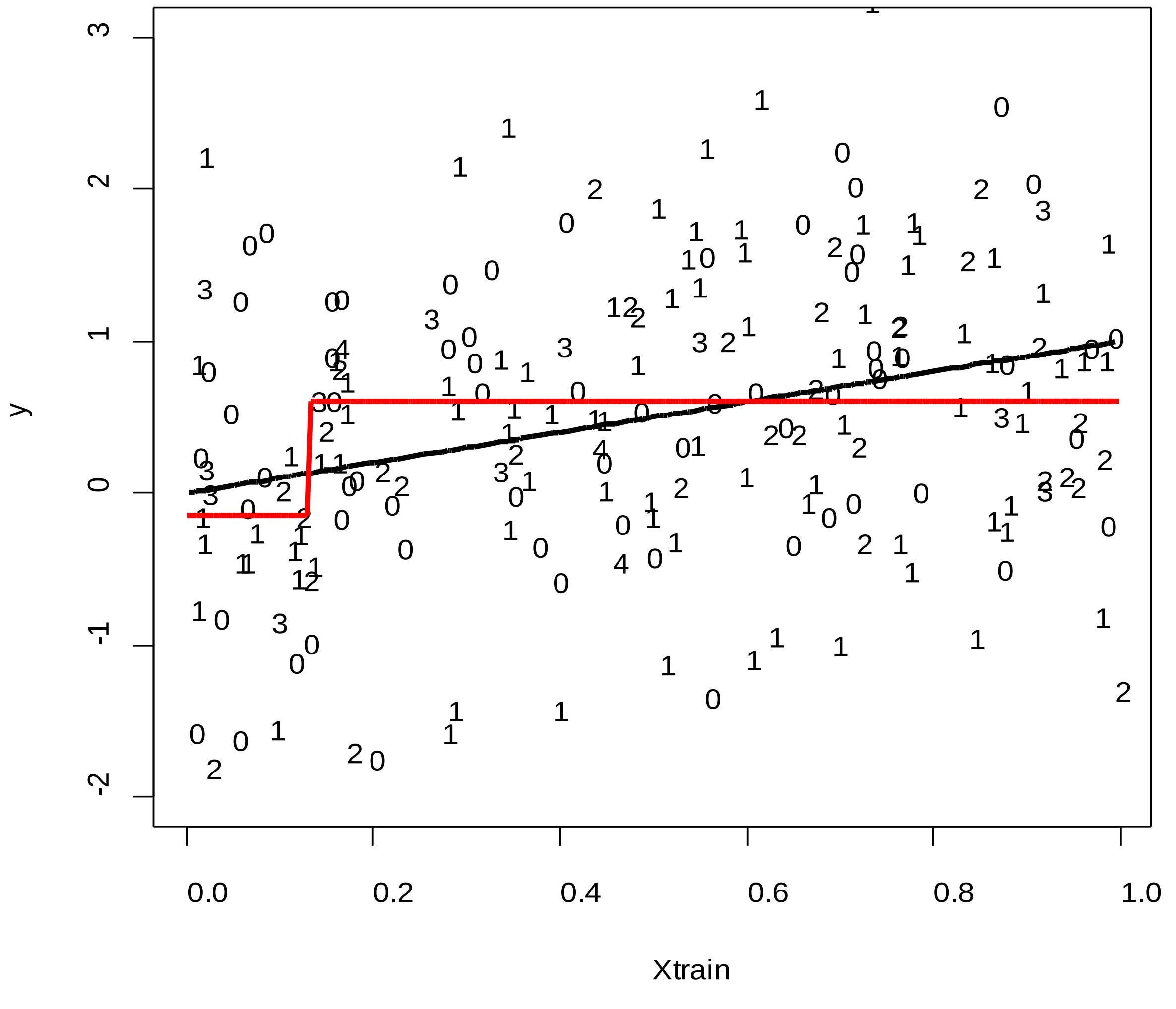


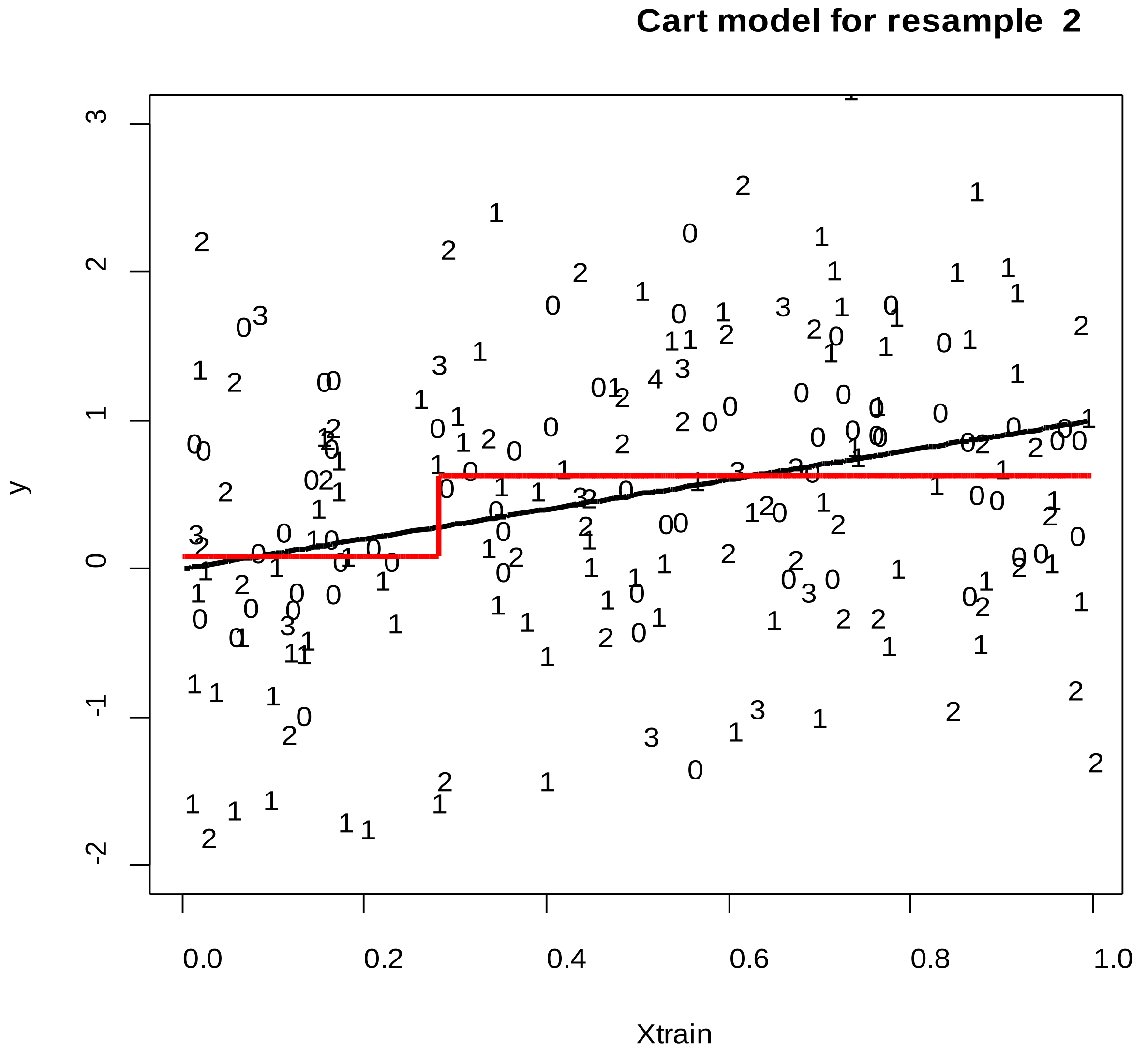
Training sample and true regressi
3
2
1
0
-1
-2
y
0.0
0.2
0.4
0.6
0.8
1.0
Xtrain

Cart model for training sample

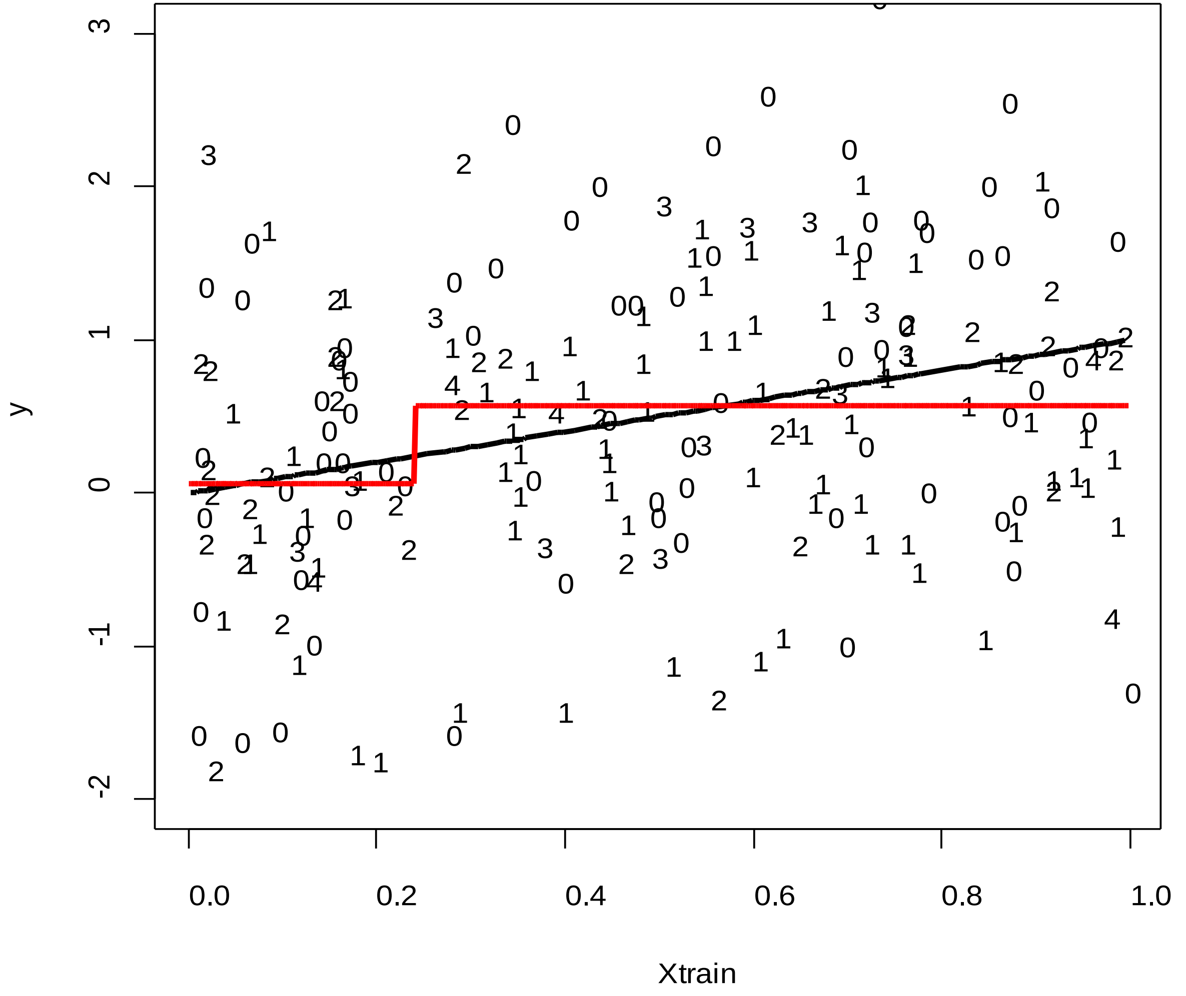


Cart model for resample 1

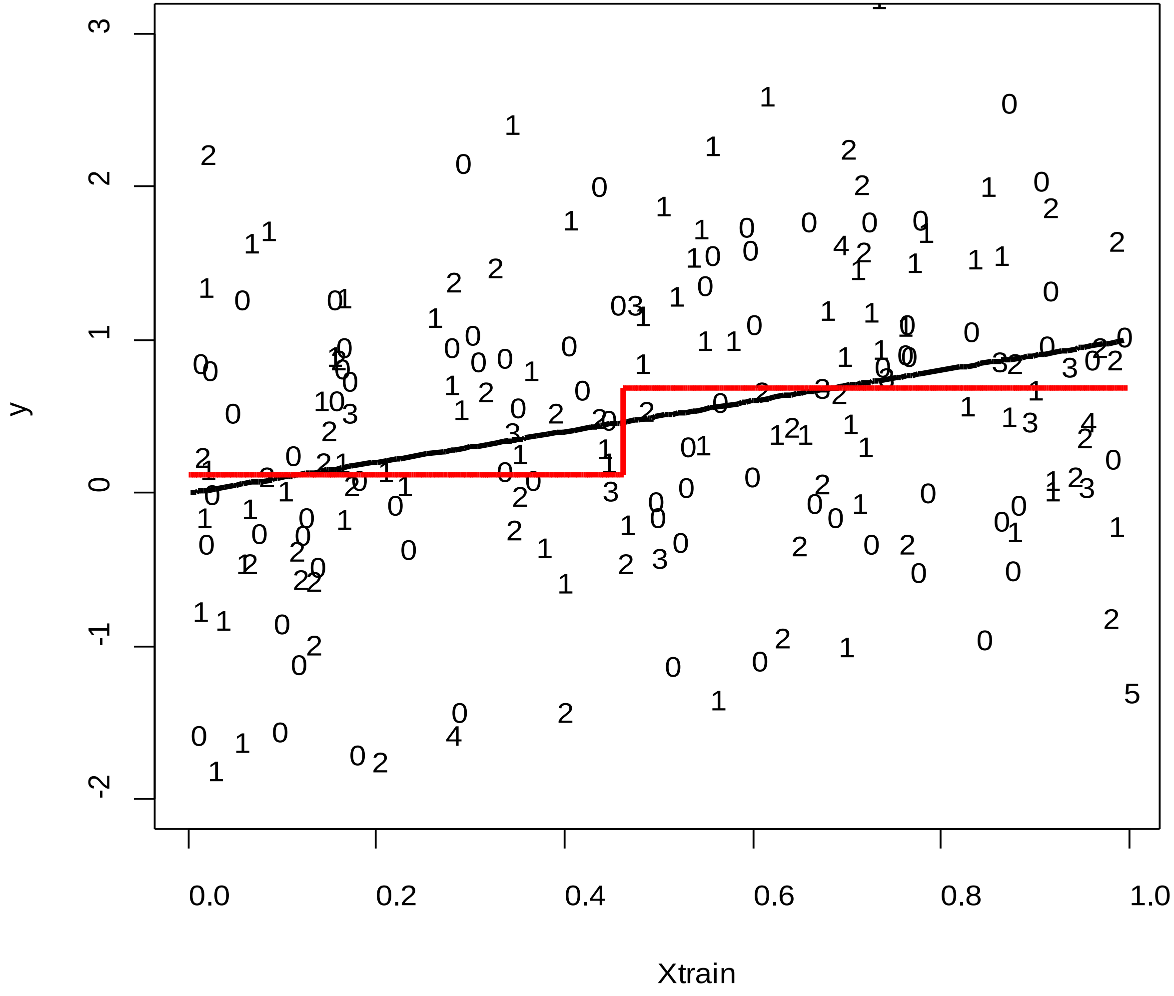



Cart model for resample 2
y
Xtrain
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
3

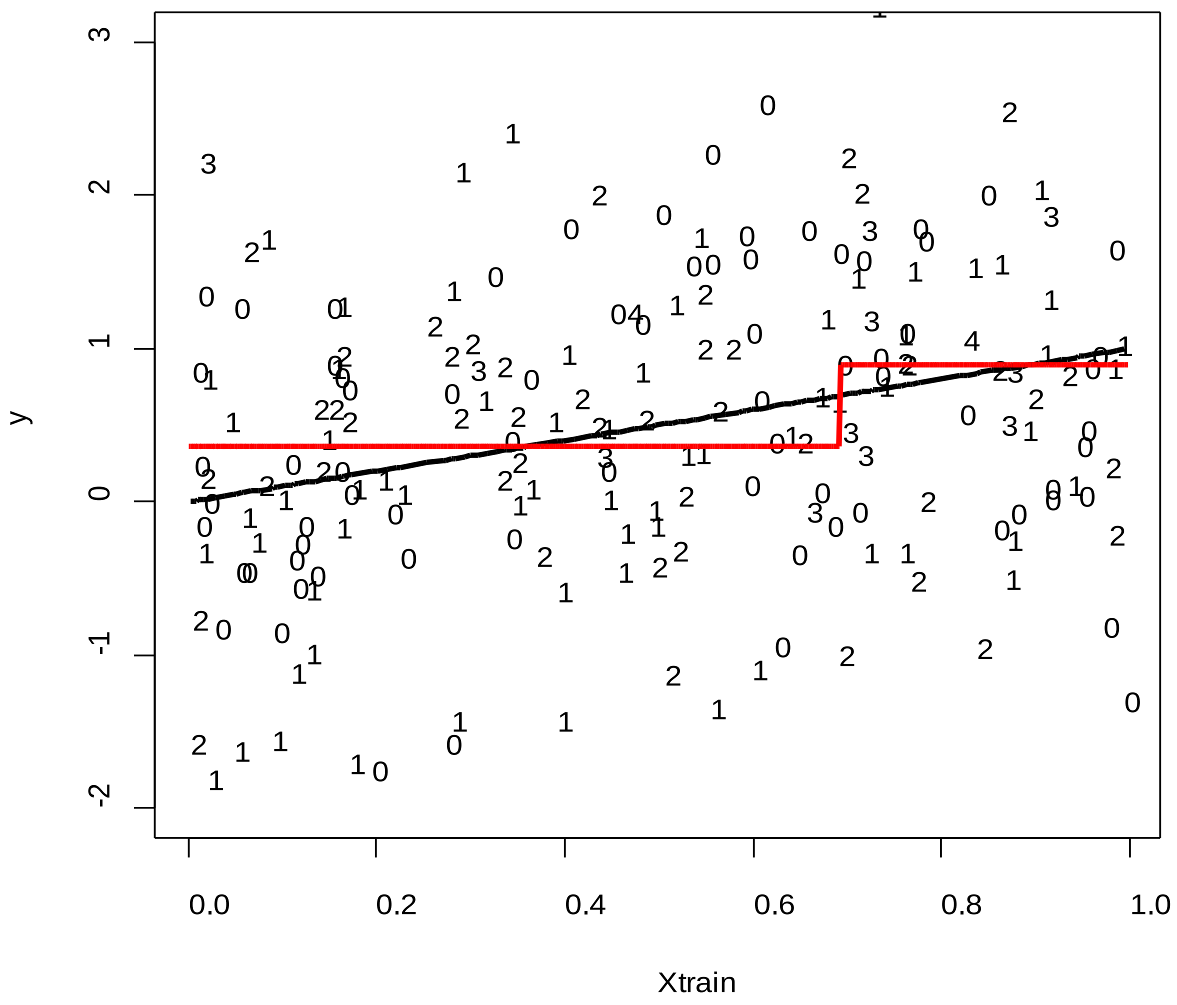
Cart model for resample 3



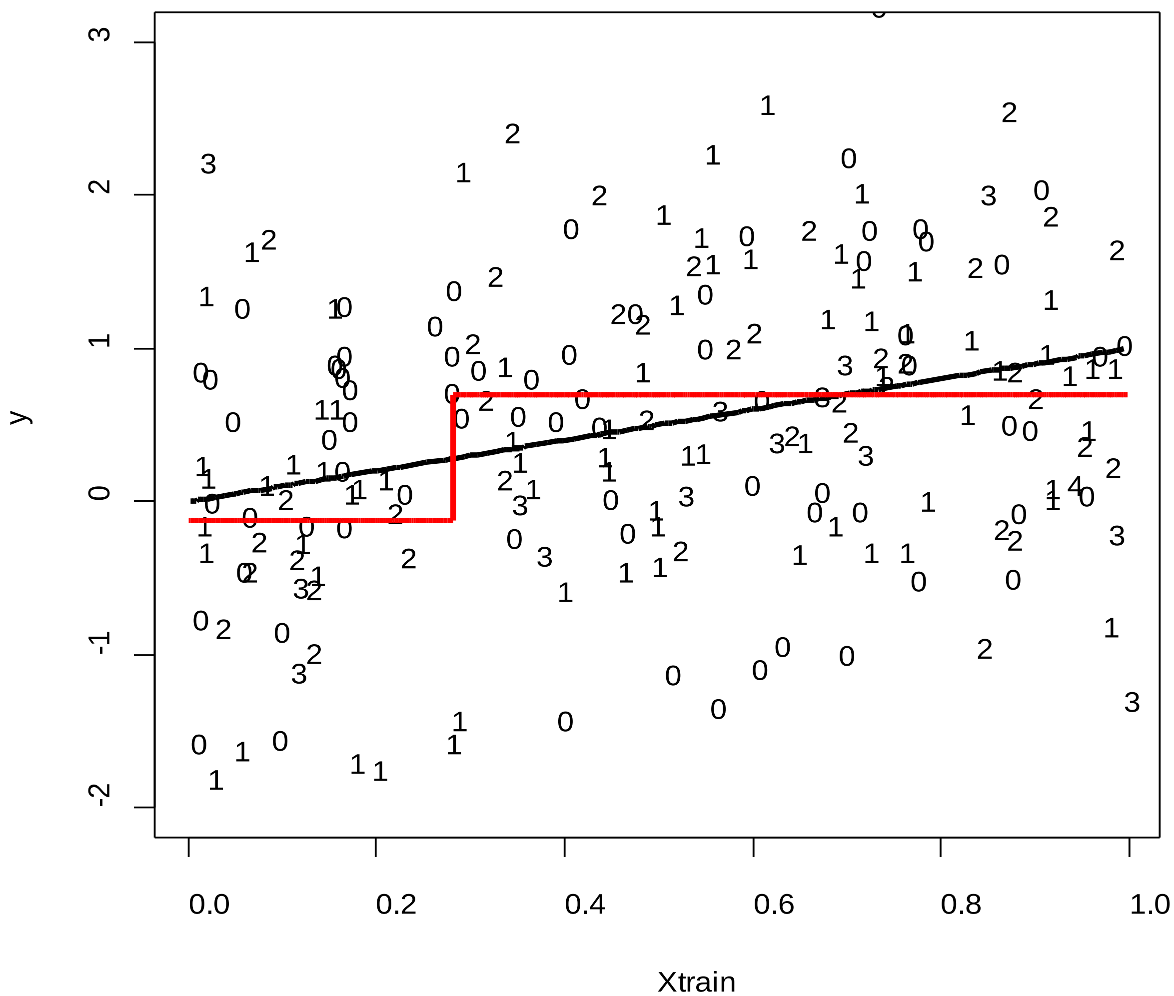
Cart model for resample 4

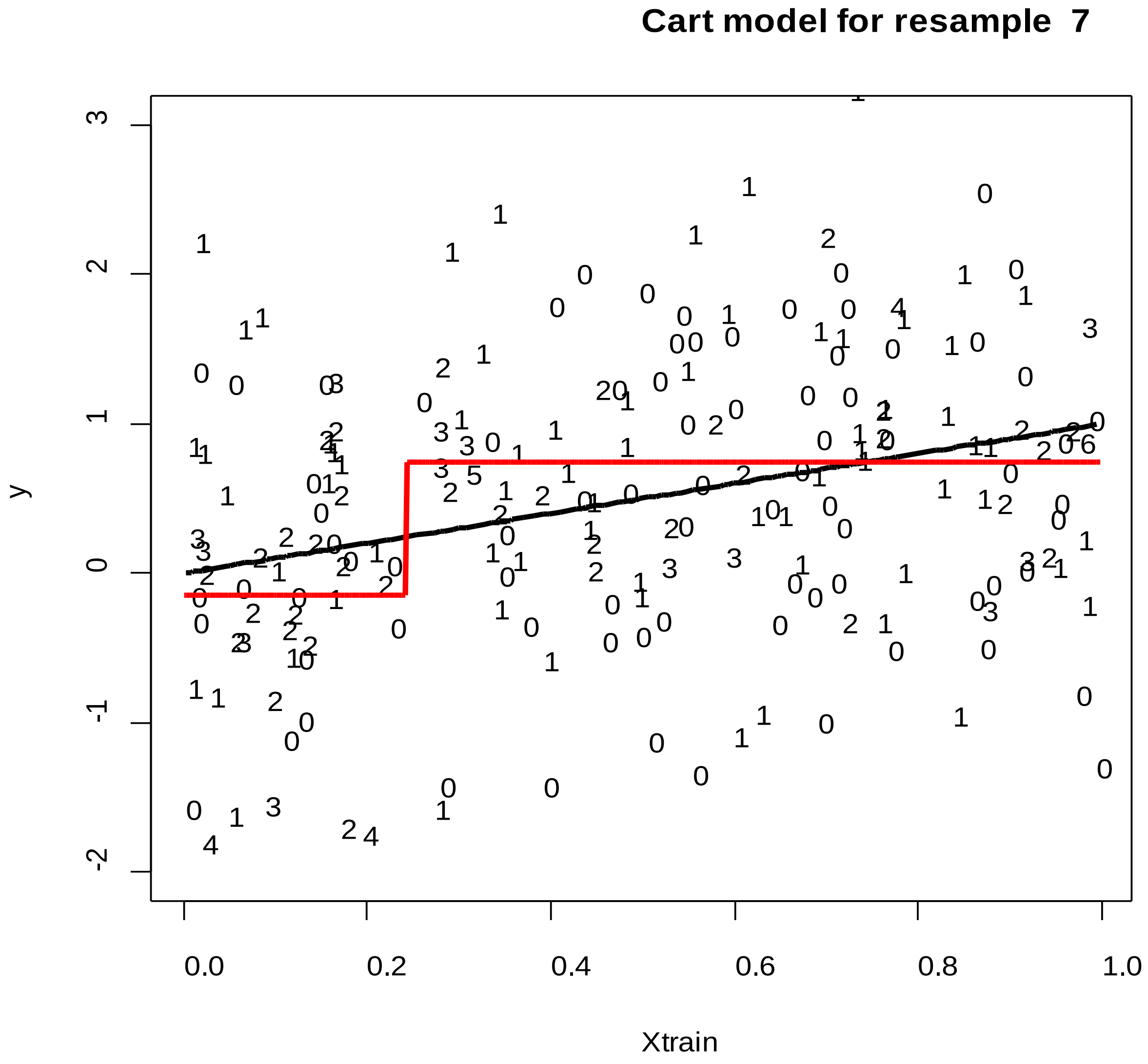


Cart model for resample 5

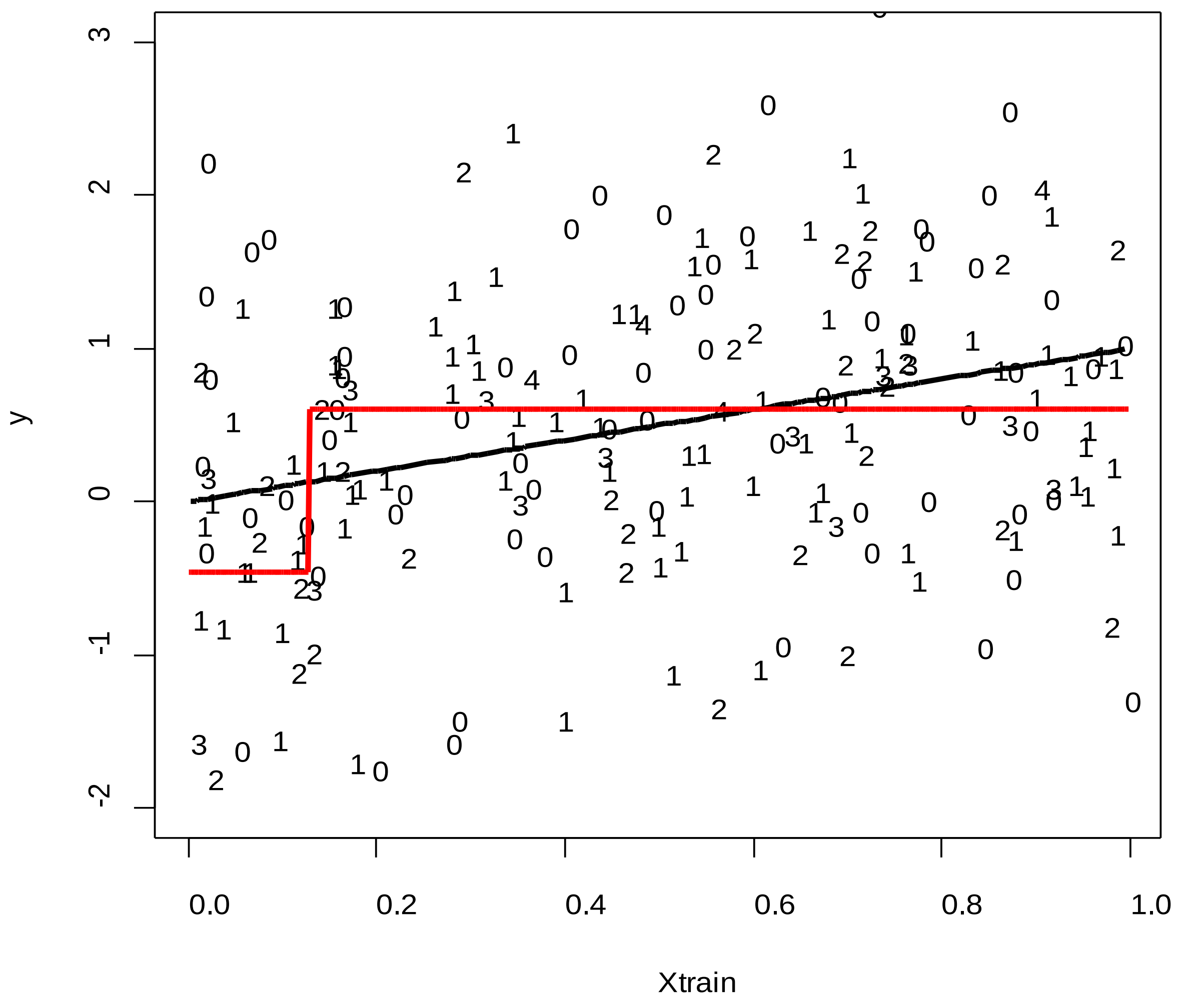


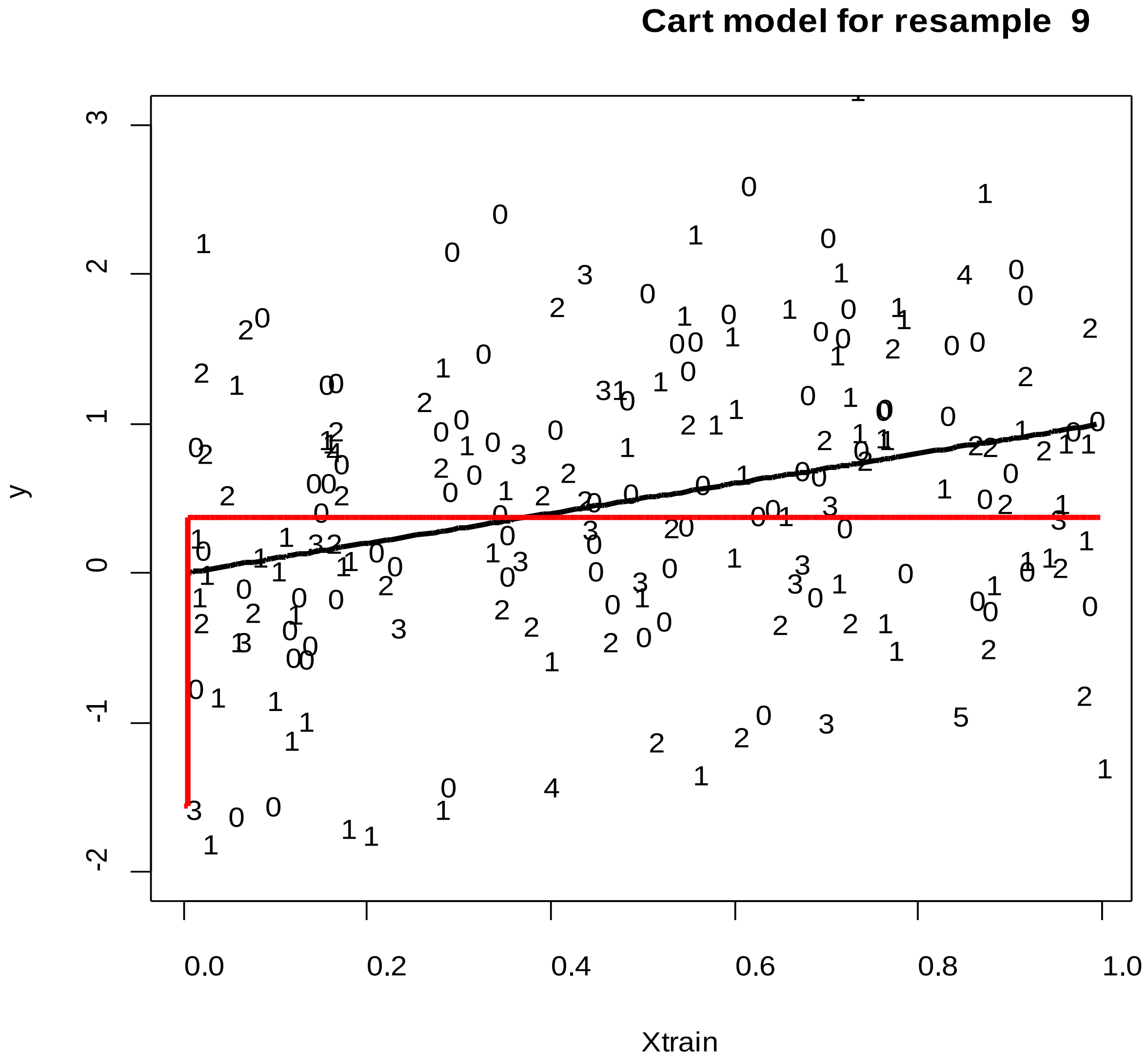
Cart model for resample 6



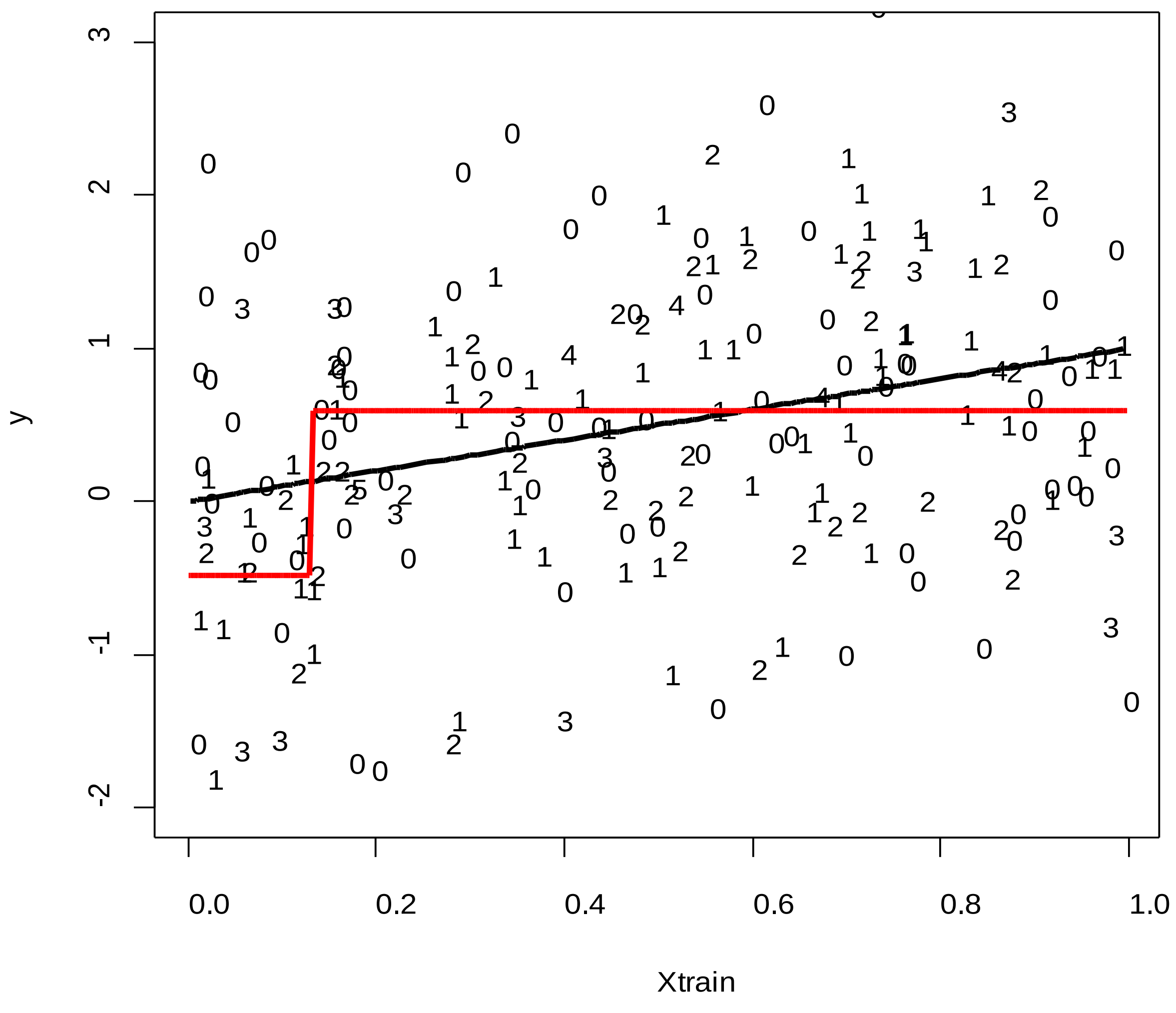

Cart model for resample 7
y
Xtrain
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
3

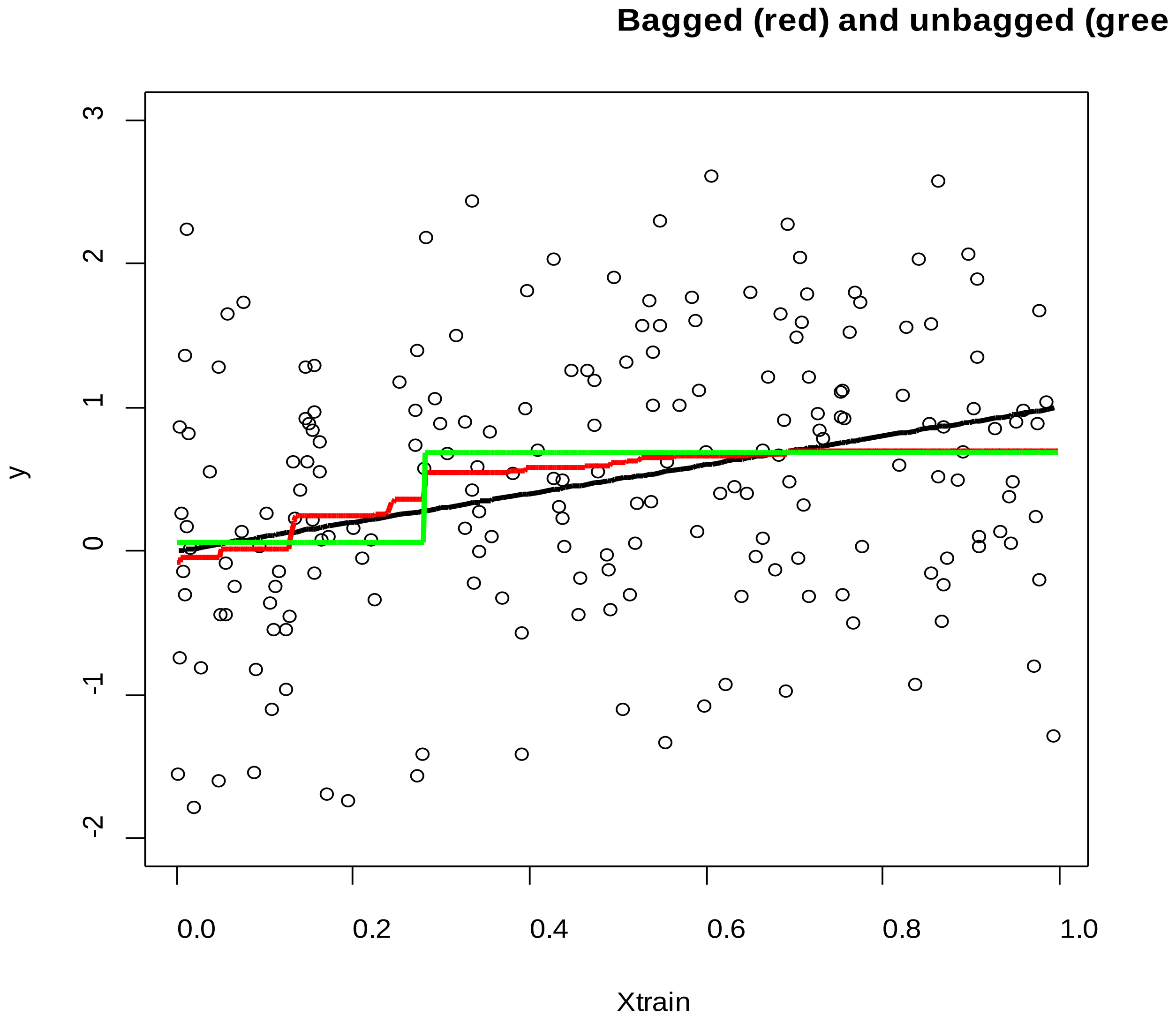
Cart model for resample 8



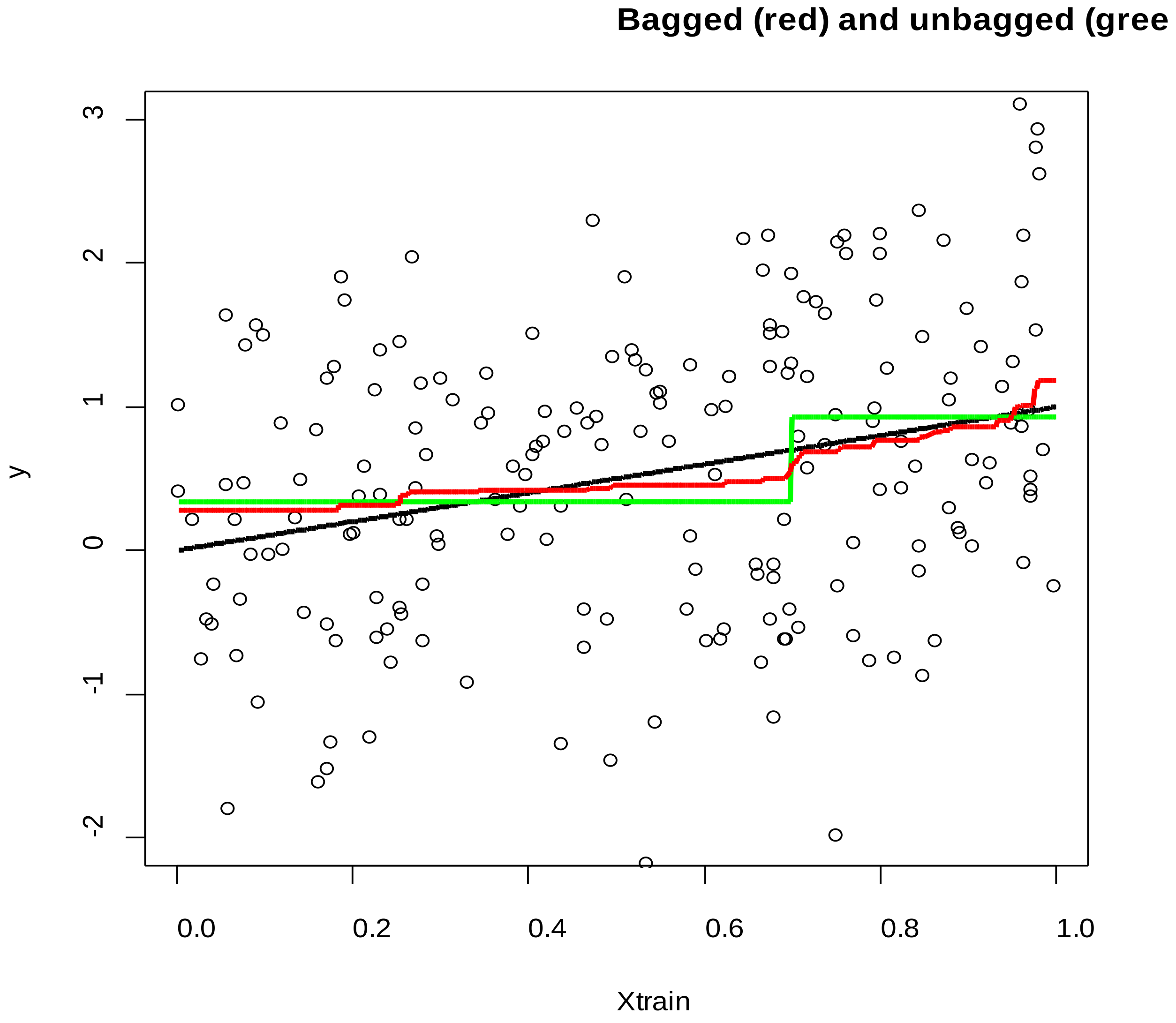

Cart model for resample 9
y
Xtrain
0.0
0.2
0.4
0.6
0.8
1.0
-2
-1
0
1
2
3

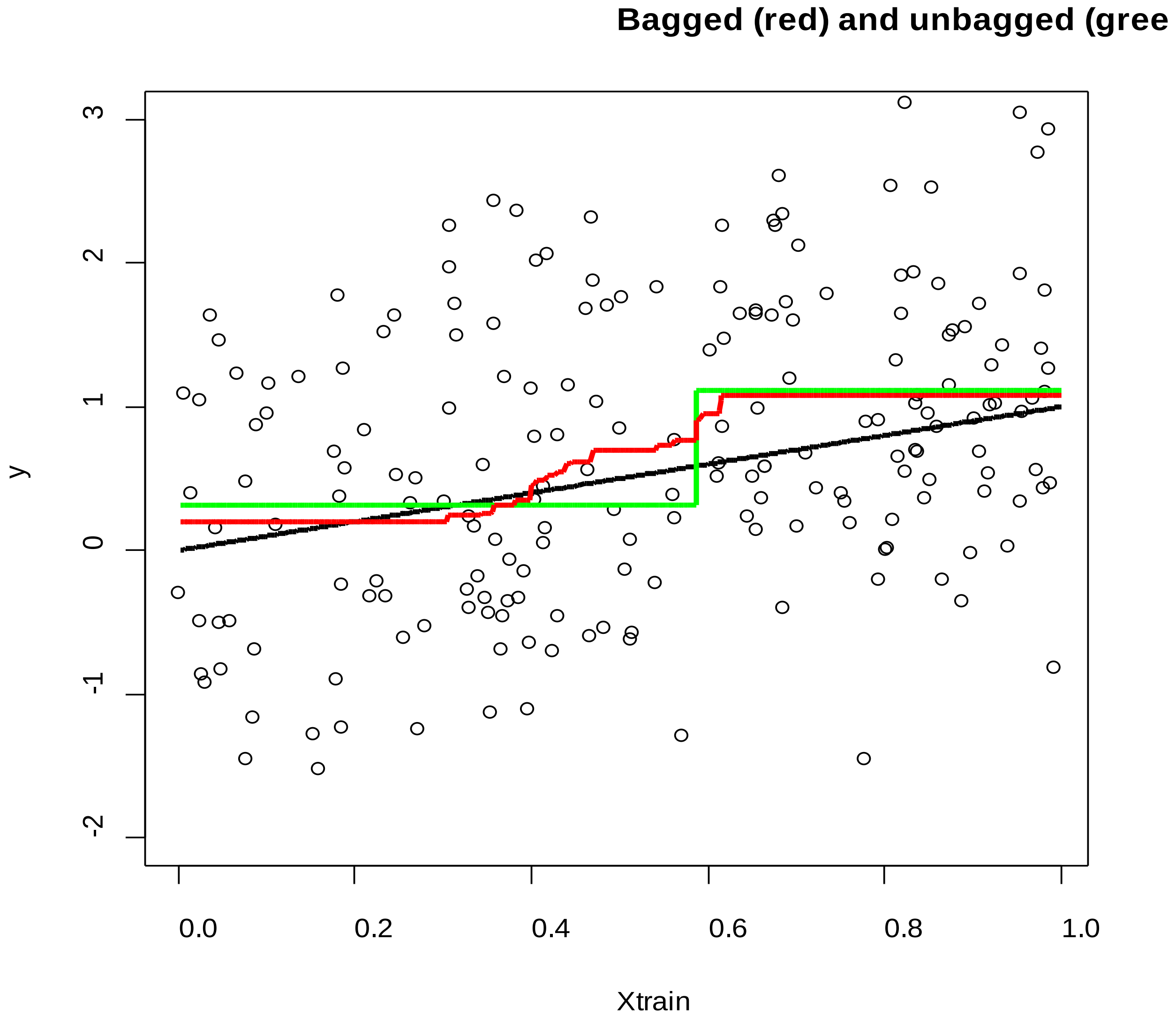
Cart model for resample 10




Bagged (red) and unbagged (gree
y
Xtrain
3
2
1
0
-1
-2
0.0
0.2
0.4
0.6
0.8
1.0

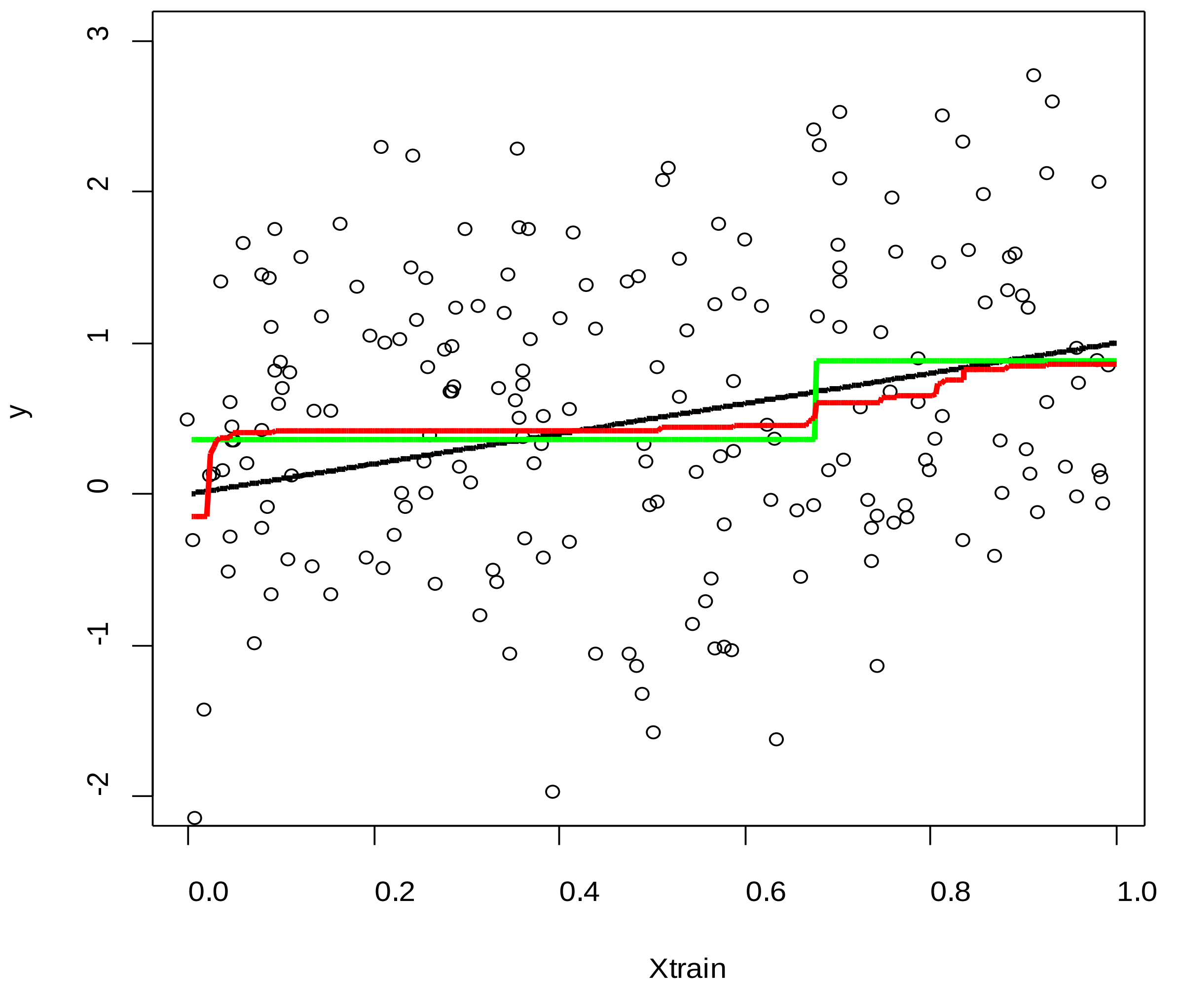
Next, compare bagged and unbagged models
for 9 more training samples.

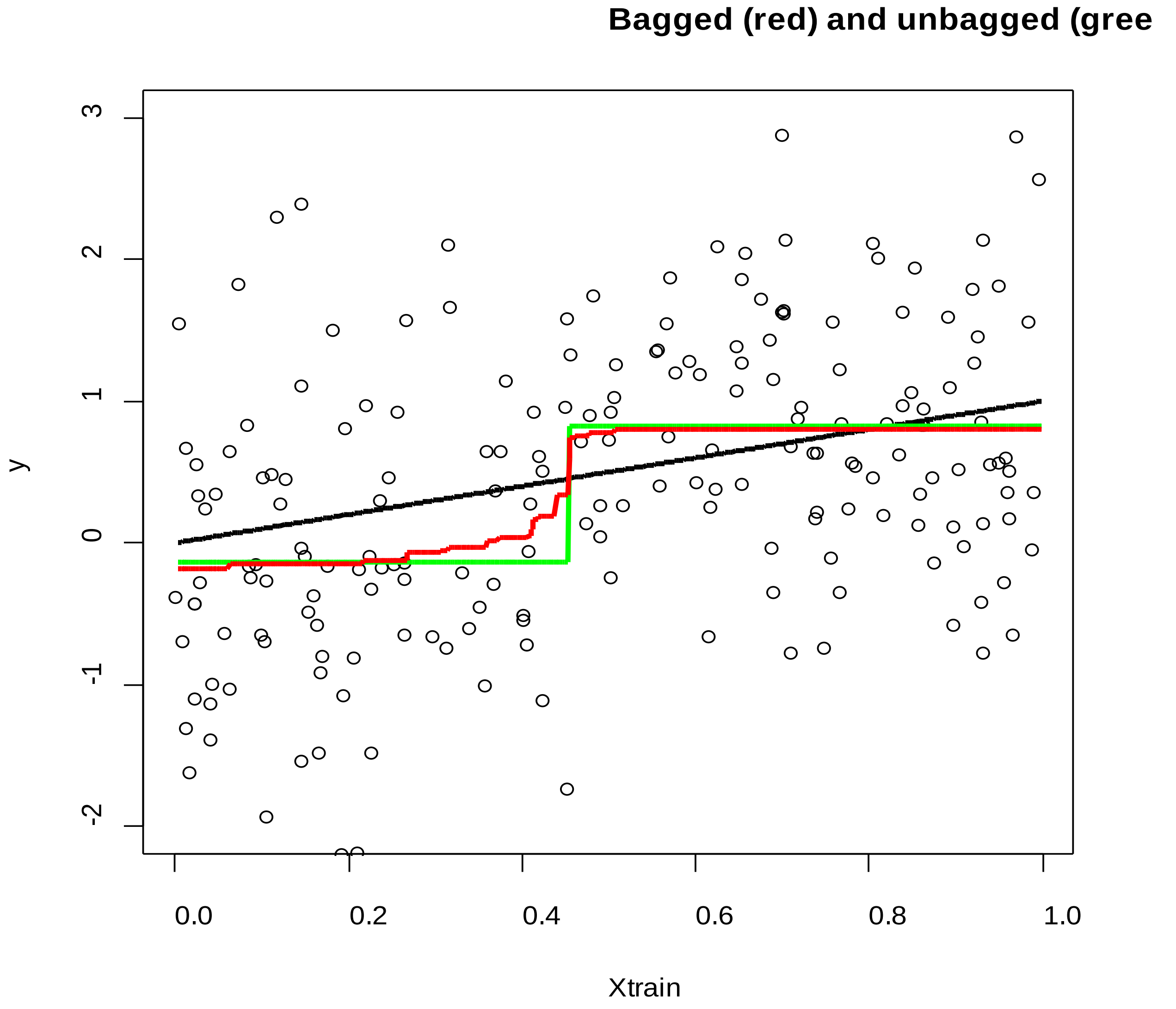

Bagged (red) and unbagged (gree
y
3
2
1
0
-1
-2
0.0
0.2
0.4
0.6
0.8
1.0
Xtrain



Bagged (red) and unbagged (gree
y
3
2
1
0
-1
-2
0.0
0.2
0.4
0.6
0.8
1.0
Xtrain

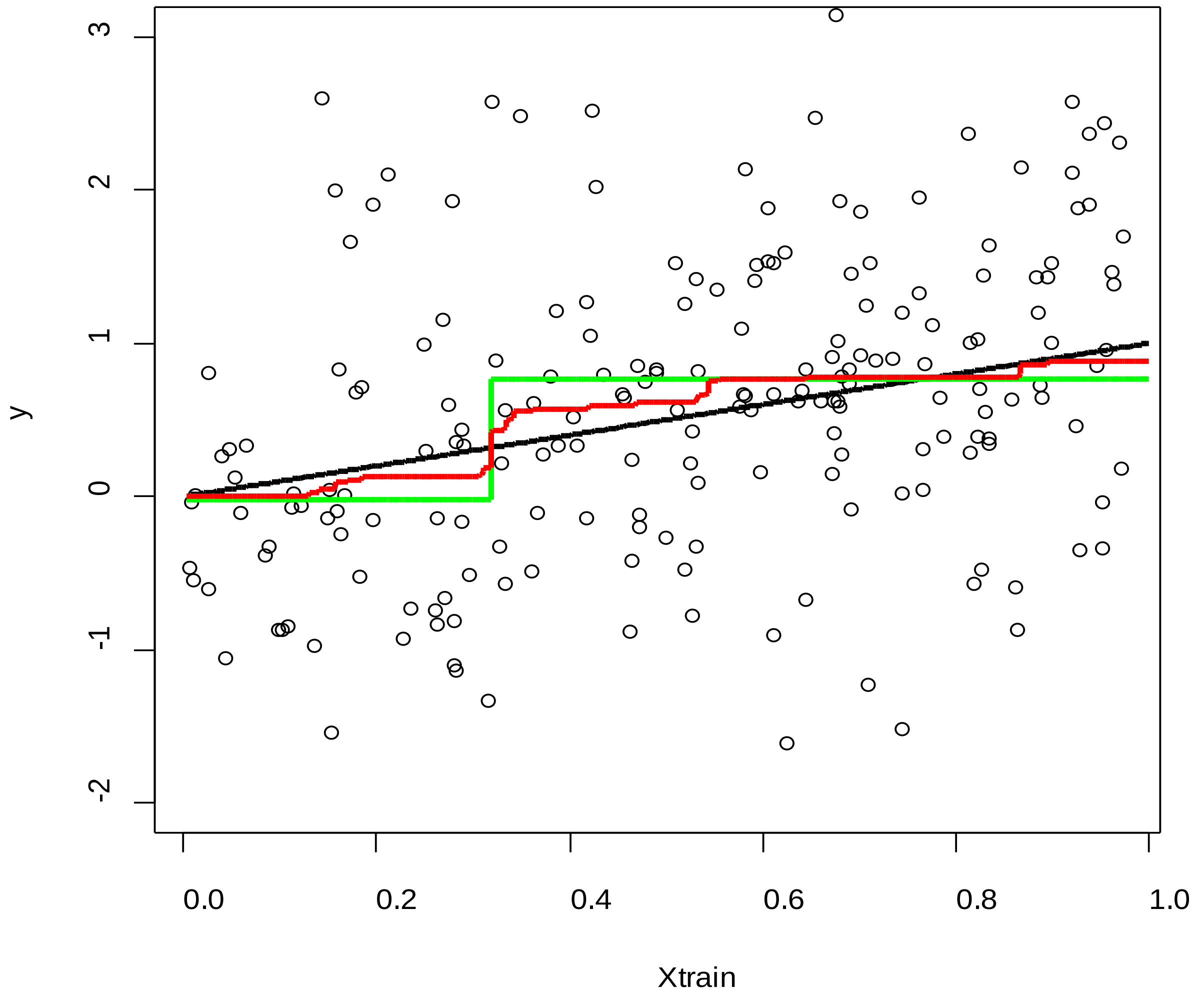
Bagged (red) and unbagged (gree

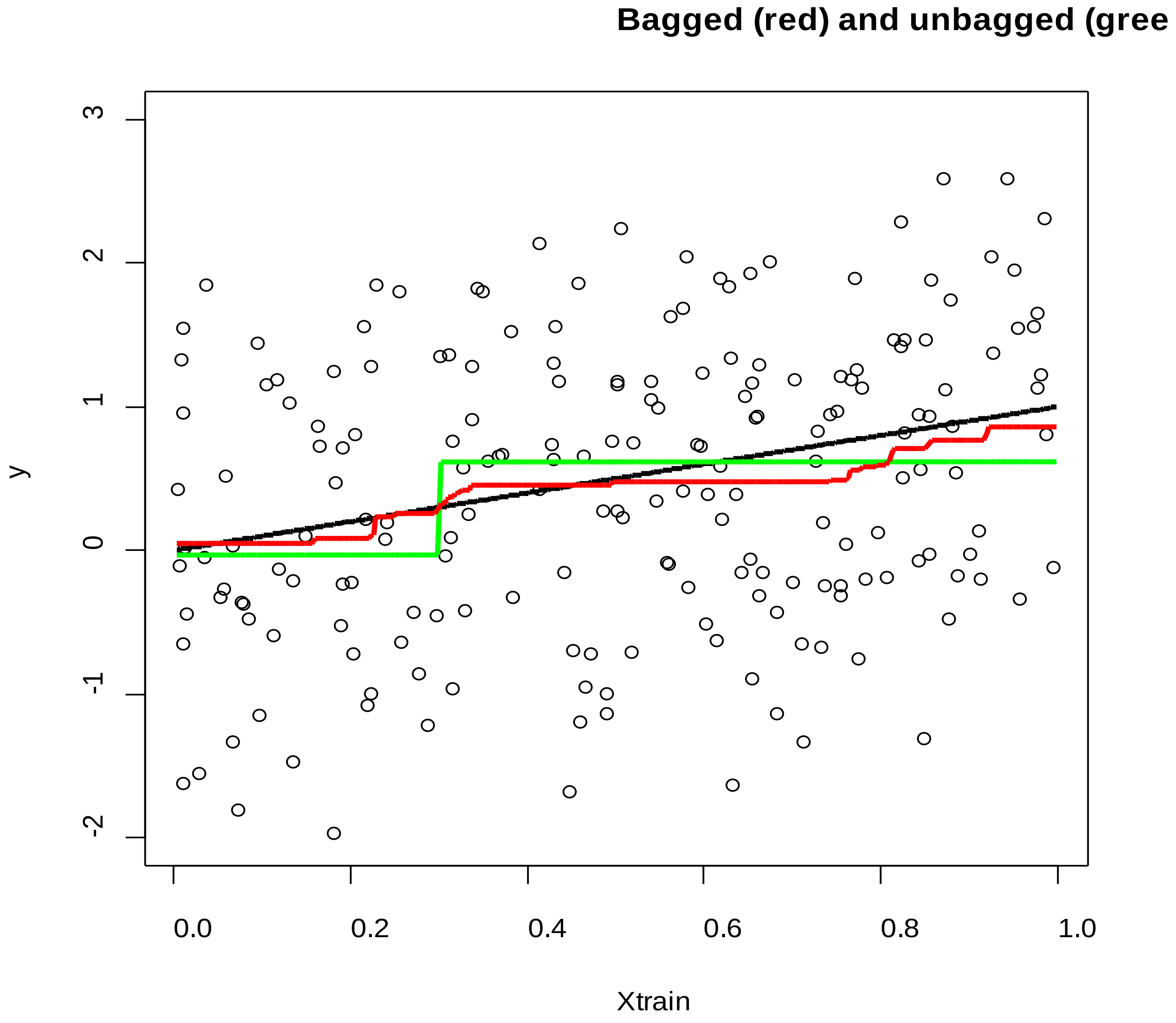


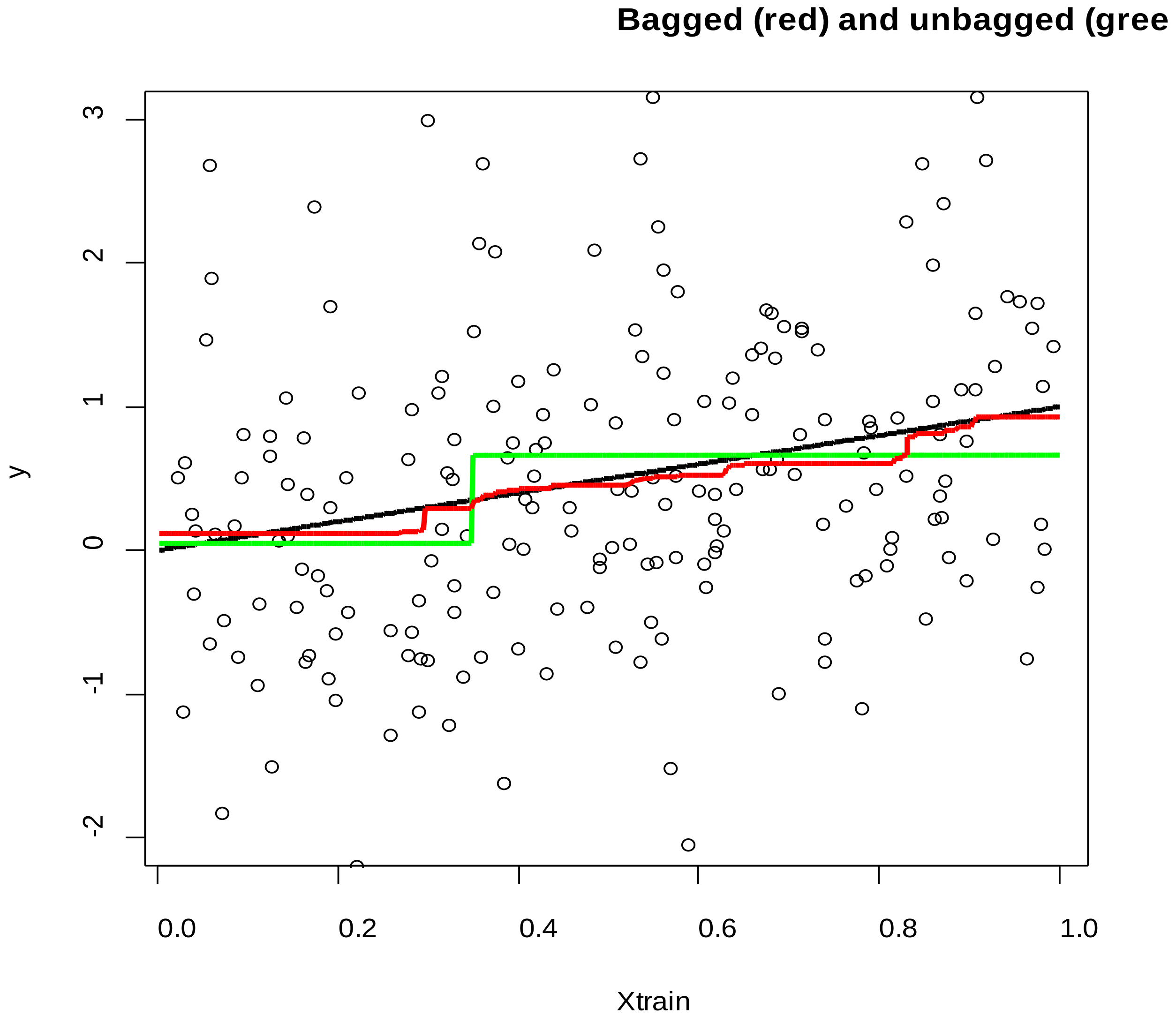


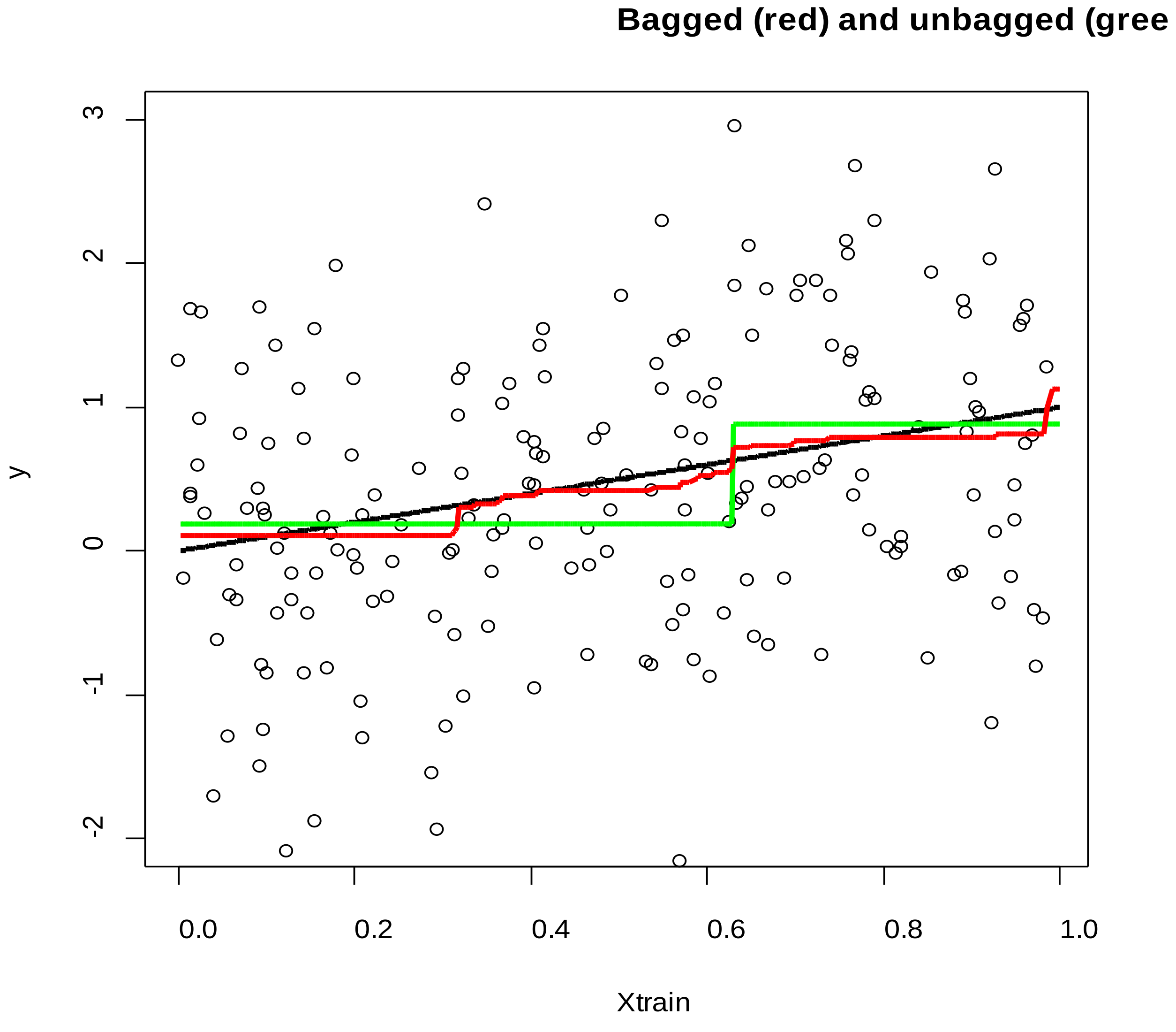
Bagged (red) and unbagged (gree
y
3
2
1
0
-1
-2
0.0
0.2
0.4
0.6
0.8
1.0
Xtrain

Bagged (red) and unbagged (gree

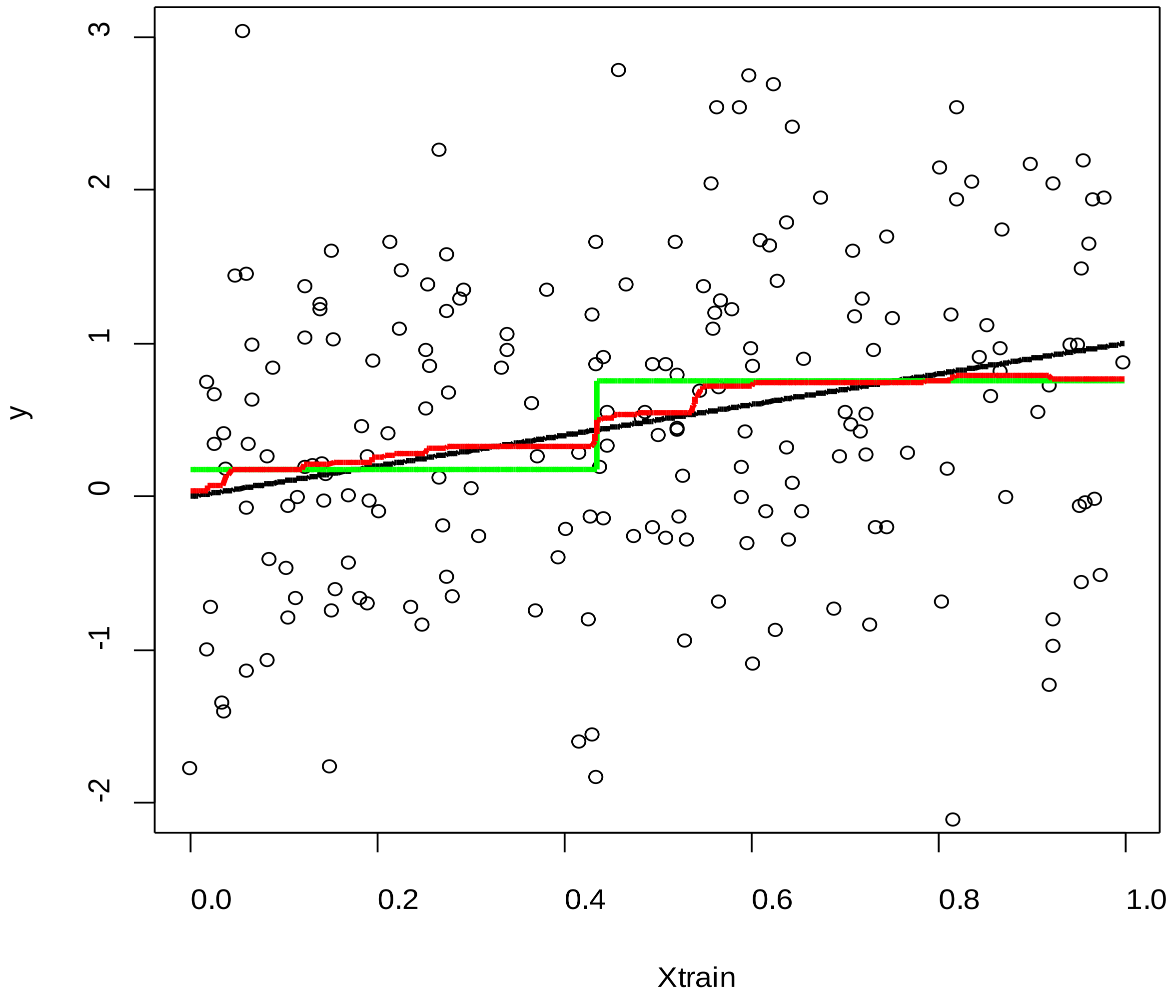



Bagged (red) and unbagged (gree
3
2
1
0
-1
-2
y
0.0
0.2
0.4
0.6
0.8
1.0
Xtrain


Bagged (red) and unbagged (gree
y
3
2
1
0
-1
-2
0.0
0.2
0.4
0.6
0.8
1.0
Xtrain


Bagged (red) and unbagged (gree
y
3
2
1
0
-1
-2
0.0
0.2
0.4
0.6
0.8
1.0
Xtrain

Bagged (red) and unbagged (gree



Compare predictive performance of bagged and unbagged models

Let $f(x) = \mathbf{E}(Y \mid x)$ be the true regression function, and let $\sigma^2(x)$ be the conditional variance of Y at x. Then

$$\mathbf{E}_Y \mathbf{E}_{\mathcal{X}}(Y(x) - p(x;\mathcal{X}))^2 = \sigma^2(x) + \mathbf{E}_{\mathcal{X}}(p(x;\mathcal{X}) - f(x))^2$$

Expected squared prediction error(x) =

conditional variance(x) +

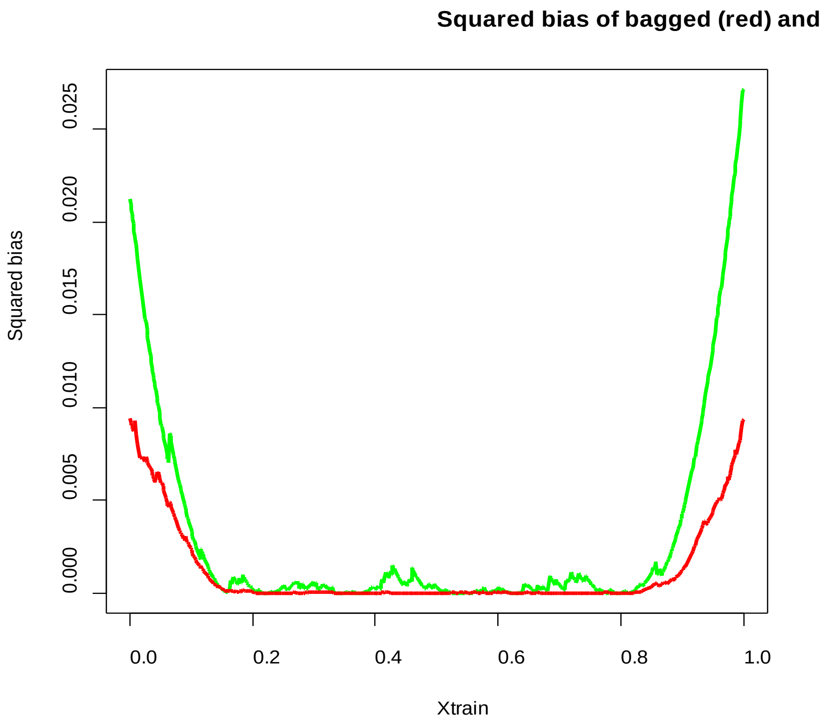
expected squared estimation error(x)

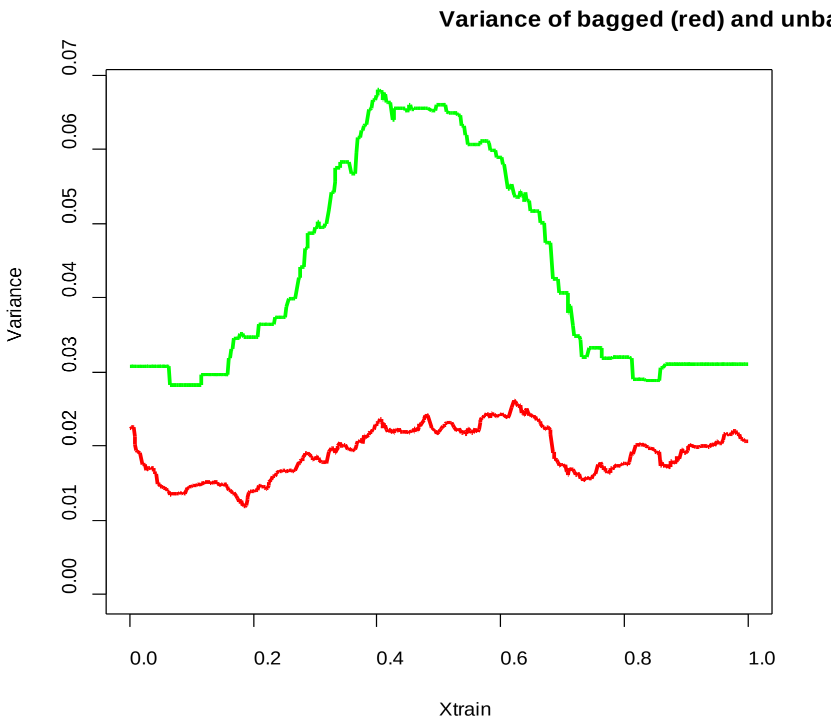
$$\mathbf{E}_{\mathcal{X}}(p(x;\mathcal{X}) - f(x))^2 = \mathbf{V}_{\mathcal{X}} p(x;\mathcal{X}) + (\mathbf{E}_{\mathcal{X}} p(x;\mathcal{X}) - f(x))^2$$

Expected squared estimation error(x) =

variance of model(x) +

squared bias of model(x)





In this example, bagged model has smaller bias *and* smaller variance than unbagged model

Breiman's heuristic

Recall the formula for the expected squared prediction error:

$$\mathbf{E}_Y \mathbf{E}_{\mathcal{X}} (Y(x) - p(x; \mathcal{X}))^2 = \sigma^2(x) + \mathbf{V}_{\mathcal{X}} p(x; \mathcal{X}) + (\mathbf{E}_{\mathcal{X}} p(x; \mathcal{X}) - f(x))^2$$

Suppose there was a "good fairy" giving us training samples $\mathcal{X}_1, \ldots, \mathcal{X}_m$ instead of a single training sample $\mathcal{X}$.

We then could construct models $p(x; \mathcal{X}_1), \ldots, p(x; \mathcal{X}_m)$ and average them, obtaining

$$\bar{p}(x) = \mathbf{ave}\,(p(x; \mathcal{X}_1), \ldots, p(x; \mathcal{X}_m))\,.$$

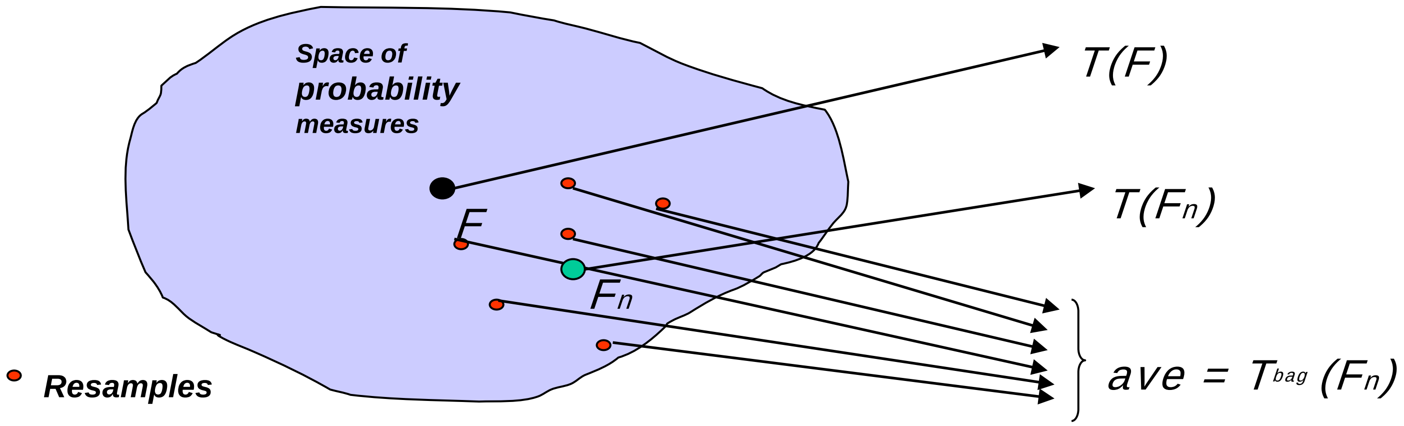
Obviously

$$\mathbf{V}\,\bar{p}(x) = \frac{1}{m} \mathbf{V}_{\mathcal{X}} p(x; \mathcal{X}_1)\,.$$

There is no "good fairy", so use Bootstrap resamples instead of new samples.

Generalizations

1. Choose resample size m different from original sample size n.



T: Functional; F: unknown distribution giving rise to observations
F_n: empirical distribution of observations

Standard approach: Estimate $T(F)$ by $T(F_n)$
Bagging: Estimate $T(F)$ by $T^{bag}(F_n)$ = average of T over resamples.

Heuristic: Smaller resample size $\Rightarrow$ resamples farther away from F_n
$\Rightarrow$ more averaging $\Rightarrow$ smaller variance, larger bias (??)

Generalizations continued

2. Draw resamples without replacement

Cuts computation in half.

Theoretical analysis of bagging

Consider functionals of the form

$$T(F) \;=\; \int \psi_1(x)\, \mathrm{d}F(x) + \int \psi_2(x_1, x_2)\, \mathrm{d}F(x_1)\, \mathrm{d}F(x_2) + \int \psi_2(x_1, x_2, x_3)\, \mathrm{d}F(x_1)\, \mathrm{d}F(x_2)\, \mathrm{d}F(x_3) + \cdots$$

(finitely many terms).

The obvious (substitution) estimate of $T(F)$ from a sample $x_1, \ldots, x_n$ is

$$T(F_n) = \frac{1}{n}\sum_i \psi_1(x_i) + \frac{1}{n^2}\sum_{ij} \psi_2(x_i, x_j) + \frac{1}{n^3}\sum_{ijk} \psi_3(x_i, x_j, x_k) + \cdots$$

Motivation

- Many statistics can be well approximated by expansions of this form.
- Can explicitly write down bagged version of T

Bagging $T(F_n)$

Let $W_1, \ldots, W_n$ be the frequencies of $x_1, \ldots, x_n$ in a resample.

If we draw resamples of size m with replacement, then the frequency vector $\underline{W}$ has a multinomial distribution.

If we draw resamples of size m without replacement, then $\underline{W}$ has a hypergeometric distribution.

The bagged version of $T(F_n)$ is

$$\begin{aligned} T^{bag}(F_n) &= \mathbf{E}_W \Bigg(\frac{1}{m} \sum_i W_i \, \psi_1(x_i) + \frac{1}{m^2} \sum_{ij} W_i W_j \, \psi_2(x_i, x_j) \\ &\quad + \frac{1}{m^3} \sum_{ijk} W_i W_j W_k \, \psi_3(x_i, x_j, x_k) + \cdots \Bigg) \end{aligned}$$

Key fact: $T^{bag}(F_n)$ is of the same form as $T(F_n)$, just with different kernels $\psi_1, \psi_2, \ldots$.

Results

Want to compare bias and variance of $T(F_n)$ – regarded as an estimate of $T(F)$ – with bias and variance of $T^{bag}(F_n)$.

Remember: $T^{bag}(F_n)$ depends on resample size m and resampling mode (with or without replacement).

(1) The effects of bagging on squared bias and variance are of order $O(1/n^2)$ (??).

(2) Bagging always increases squared bias; squared bias increases as resample size decreases.

(3) Whether or not bagging decreases or increases the variance depends on the kernels $\psi_1, \psi_2, \ldots$.

Results (continued)

(4) For every resample size $m_{wo} = \alpha\, n$ for resampling without replacement there is a corresponding resample size $m_{wi} = \frac{\alpha}{1-\alpha} n$ for resampling with replacement that results in the same variance and squared bias up to $O(1/n^2)$

- The standard Bootstarp corresponds to half-sampling.
- There are situations where choosing $m > n$ (for resampling with replacement) or $m > n/2$ (without replacement) is beneficial.

Experimental results

$X \sim U[0,1]$
$\epsilon \sim N(0,1)$

Scenario 1: $Y = \epsilon$ (no signal)
Scenario 2: $Y = I(X > 0.5) + \epsilon_i$ (step function)
Scenario 3: $Y = X + \epsilon$ (linear function)

Cart model with 2 leaves.
Bagging with 50 resamples.

Did simulations for more complex and realistic situations (not presented here). They led to the same conclusions.

A comment on bias

In the regression context, $T(F_n)$ corresponds to the model $p(x;\mathcal{X})$ estimated from the training sample $\mathcal{X}$.

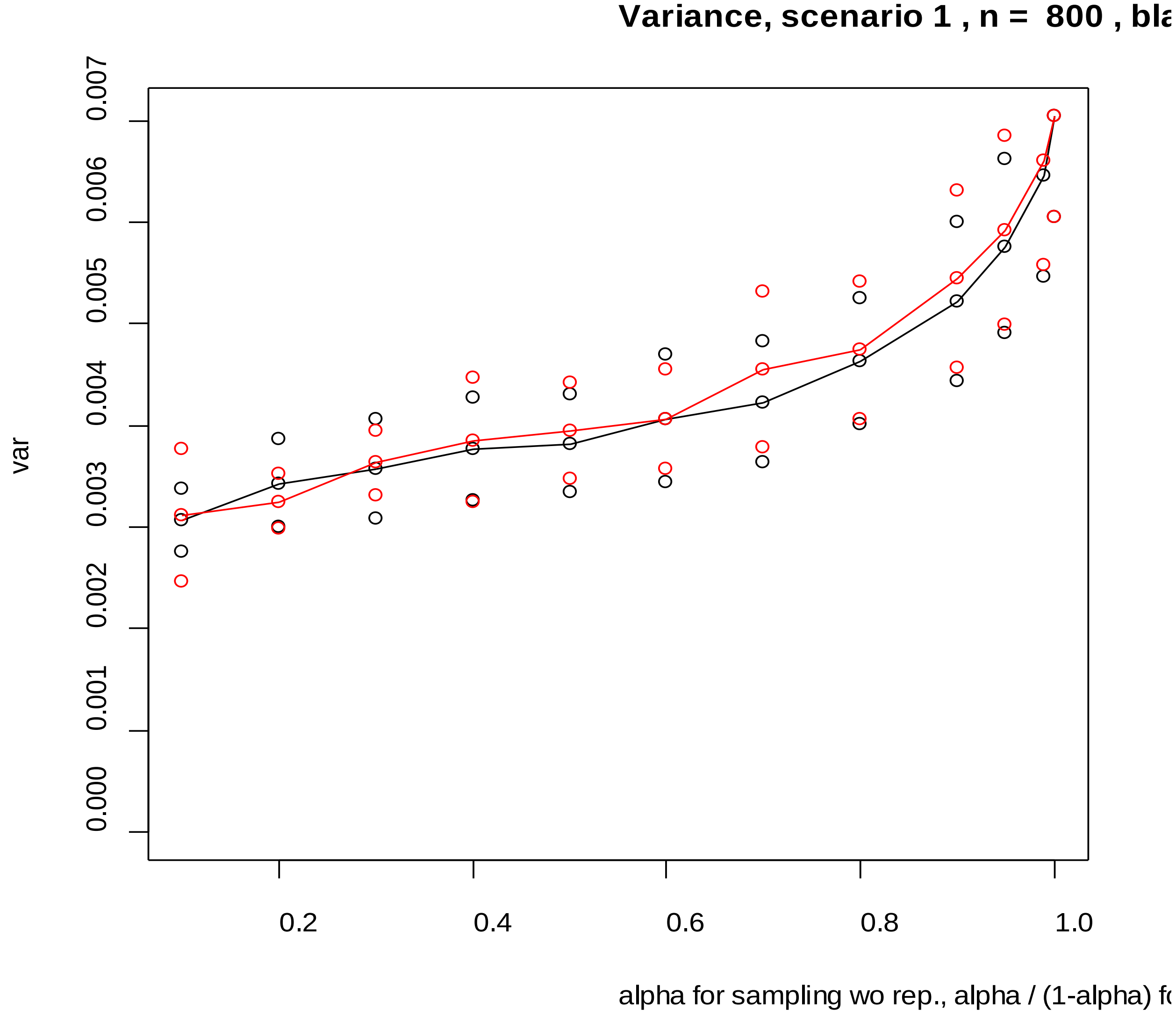
$T(F)$ corresponds to the model $p^\infty(x)$ for an infinite training sample.

In our theory, bias is defined as $\mathbf{E}\,T(F_n) - T(F) \sim \mathbf{E}_{\mathcal{X}}\,p(x;\mathcal{X}) - p^\infty(x)$

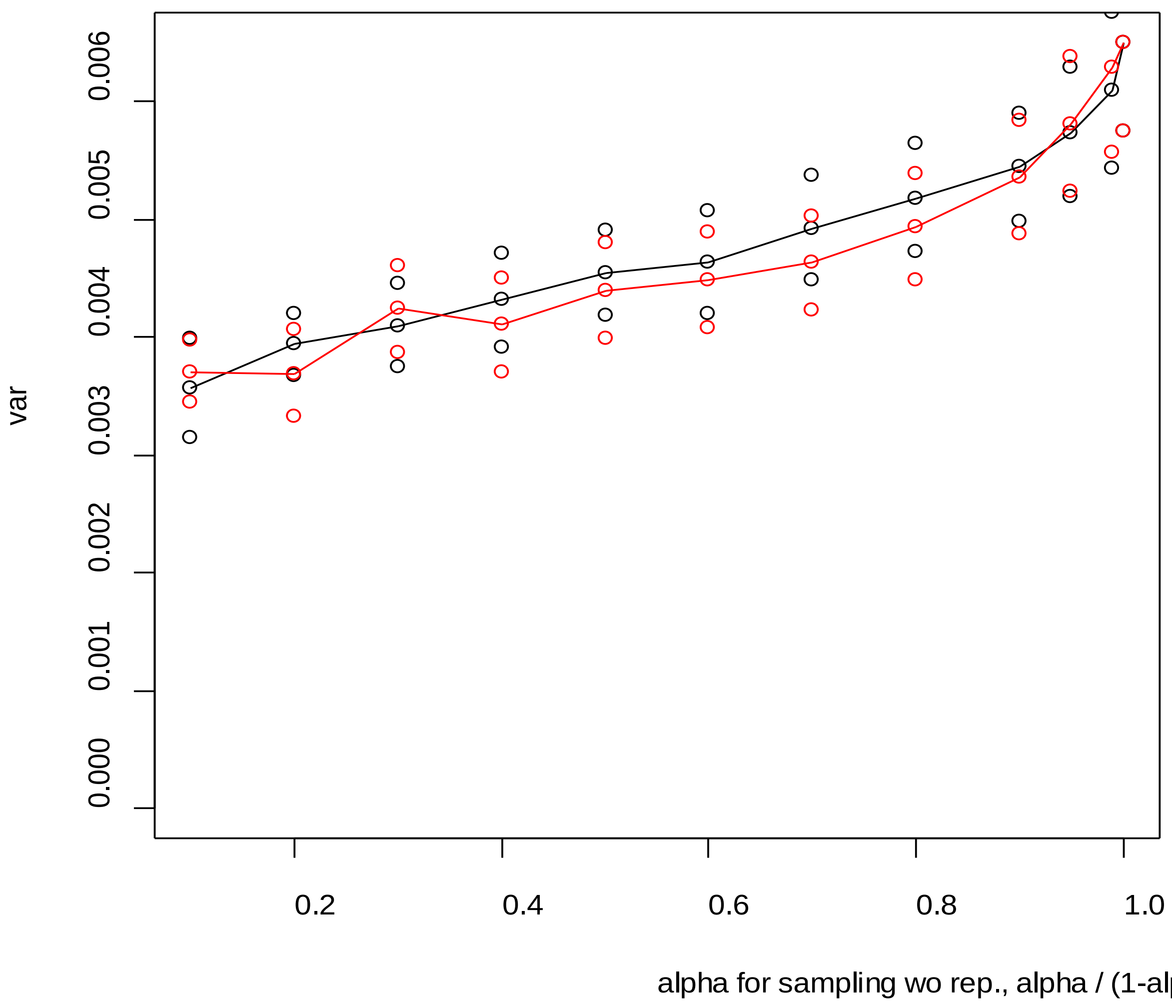
In regression analysis, bias is typically defined as $\mathbf{E}_{\mathcal{X}}\,p(x;\mathcal{X}) - f(x)$, where $f(x)$ is the true regression function.

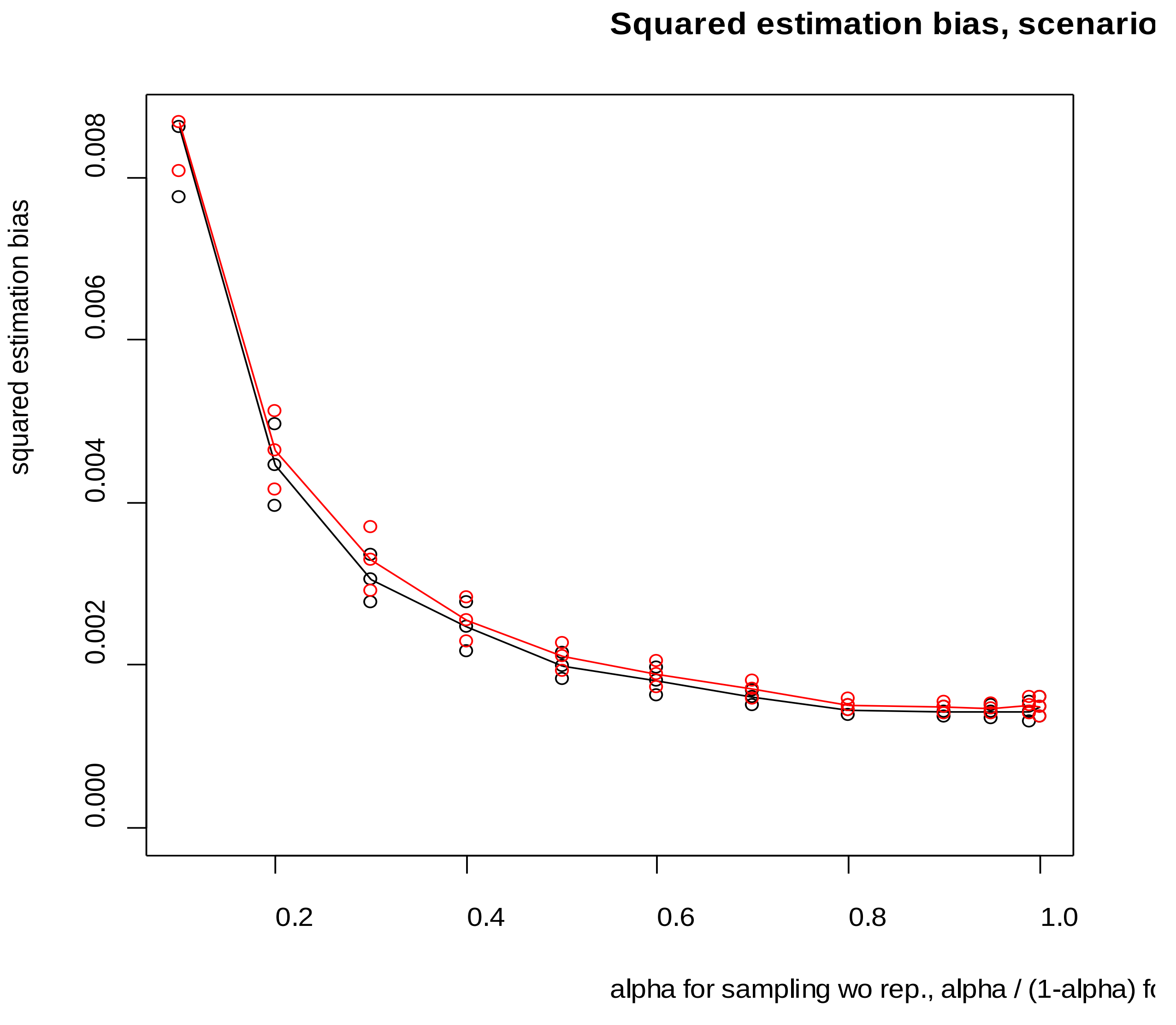
We will refer to the former as *estimation bias*.

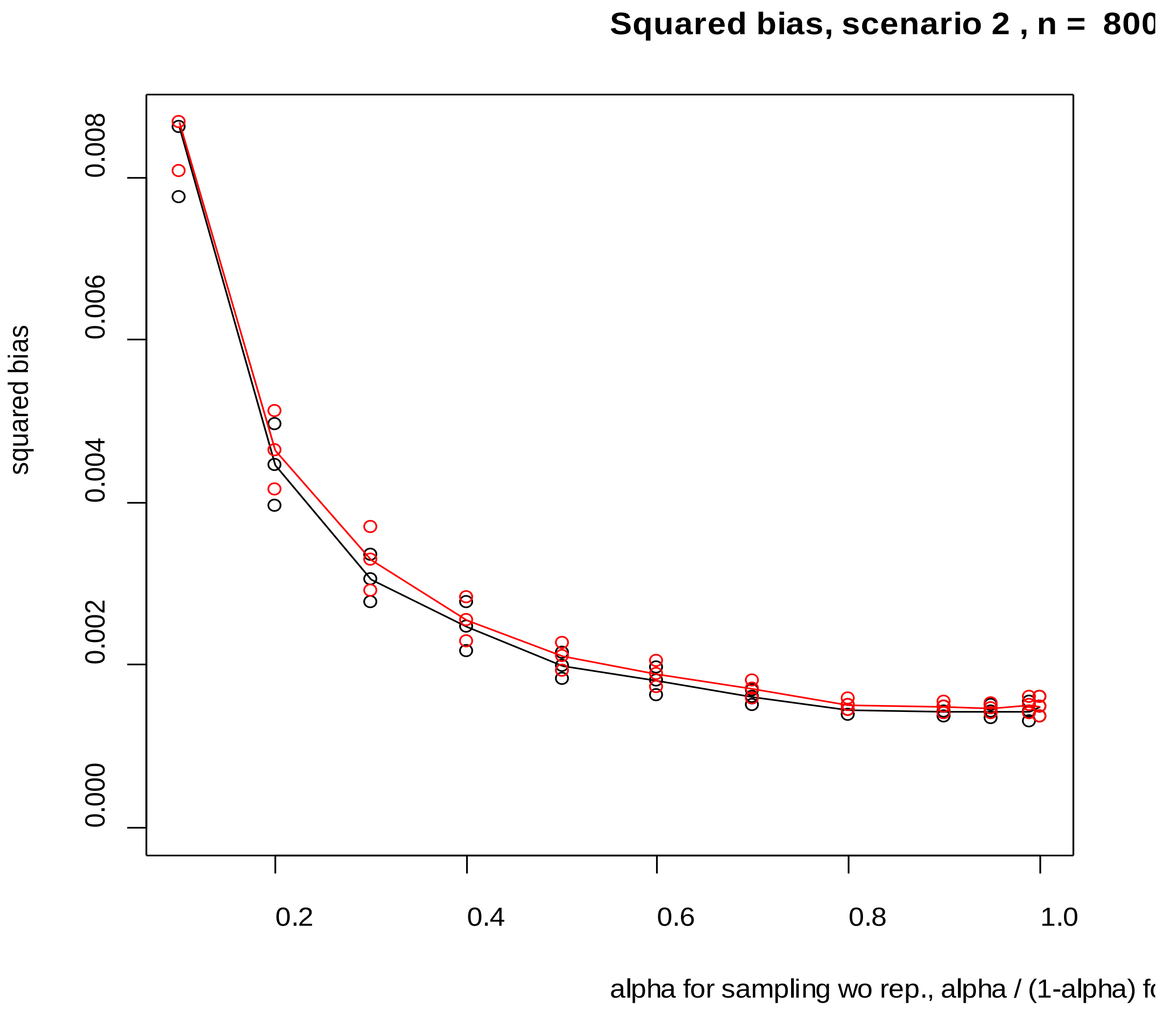
The theory predicts that estimation bias of bagged models is larger then estimation bias of unbagged model, and decreases with increasing resample size.

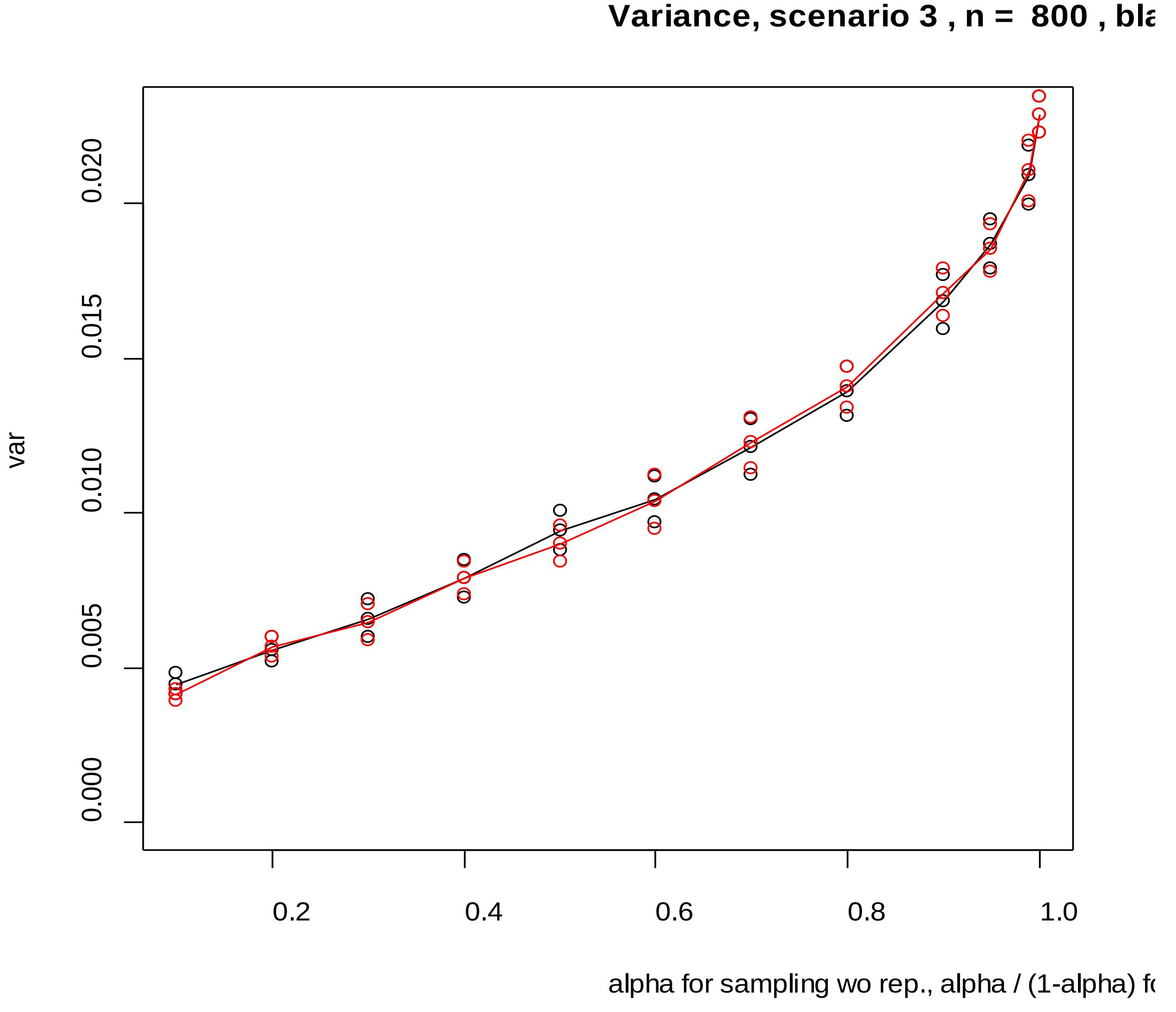

Variance, scenario 1 , n = 800 , bla
var
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.2
0.4
0.6
0.8
1.0
alpha for sampling wo rep., alpha / (1-alpha) f

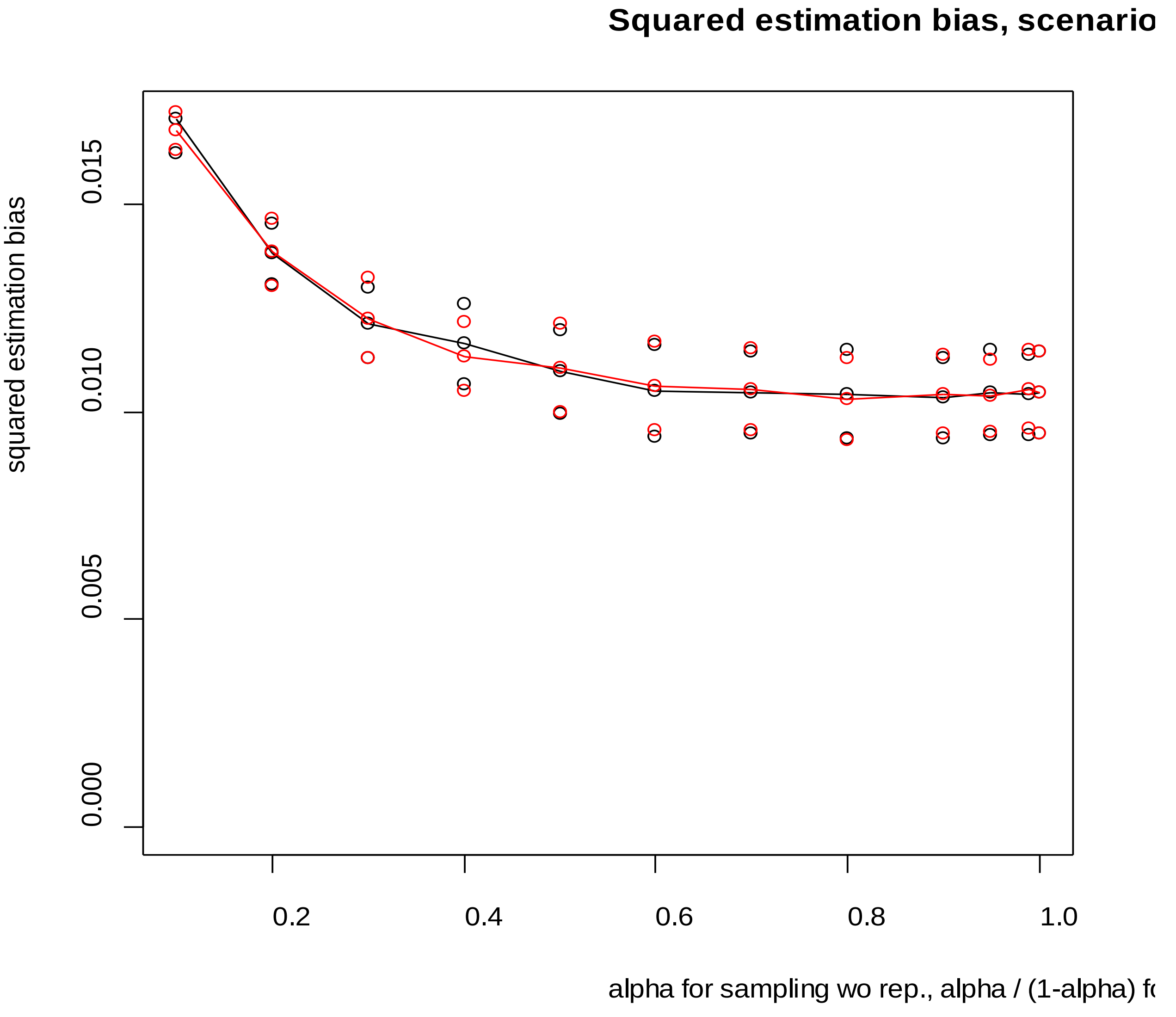
Variance, scenario 2 , n = 800 , bla

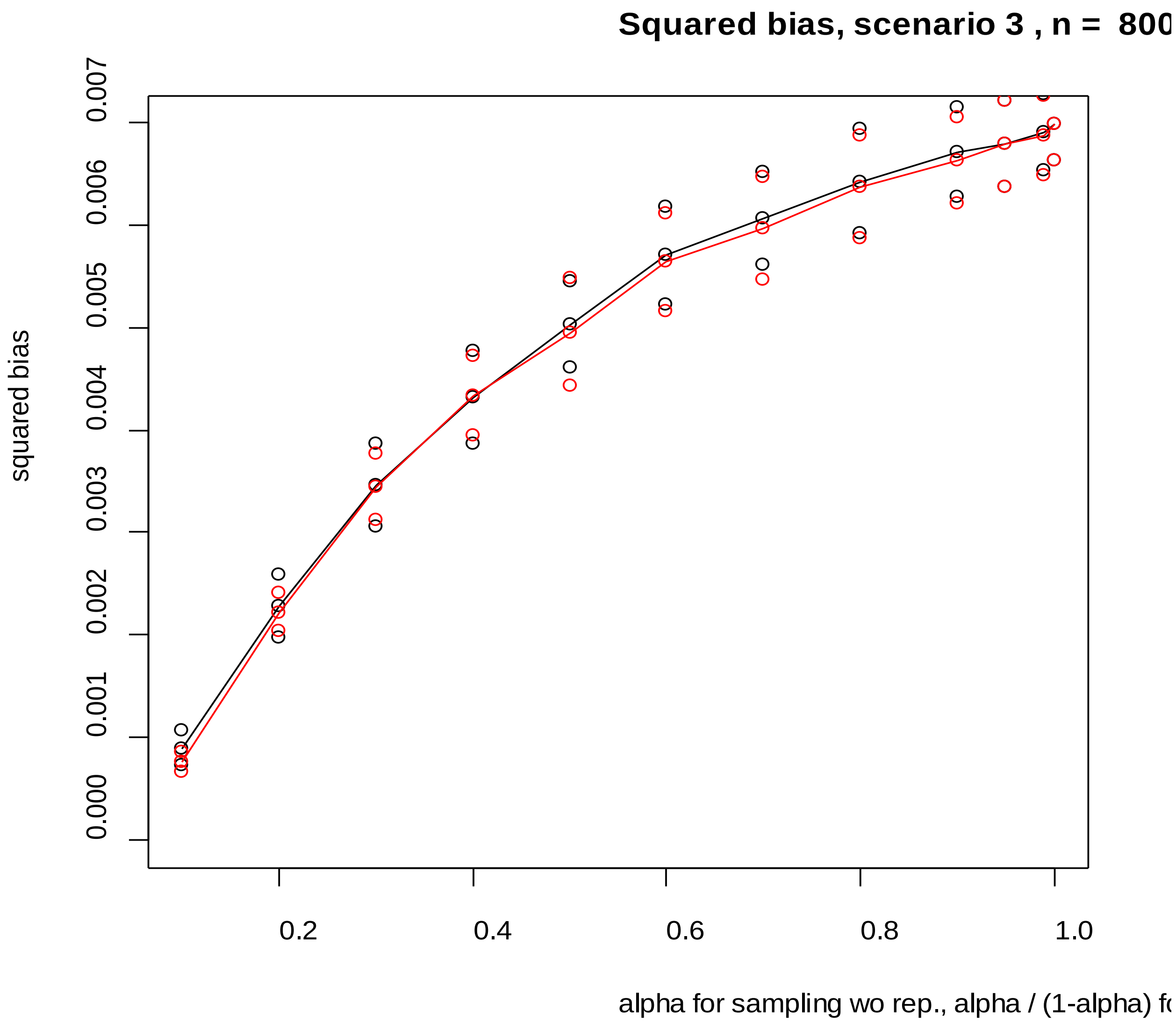



Squared estimation bias, scenario
squared estimation bias
0.000
0.002
0.004
0.006
0.008
0.2
0.4
0.6
0.8
1.0
alpha for sampling wo rep., alpha / (1-alpha) fo


Squared bias, scenario 2 , n = 800
squared bias
0.000
0.002
0.004
0.006
0.008
0.2
0.4
0.6
0.8
1.0
alpha for sampling wo rep., alpha / (1-alpha) fo


Variance, scenario 3 , n = 800 , bla
var
0.000
0.005
0.010
0.015
0.020
0.2
0.4
0.6
0.8
1.0
alpha for sampling wo rep., alpha / (1-alpha) fo


Squared estimation bias, scenario
squared estimation bias
0.000
0.005
0.010
0.015
0.2
0.4
0.6
0.8
1.0
alpha for sampling wo rep., alpha / (1-alpha) fo


Squared bias, scenario 3 , n = 800
squared bias
0.000
0.001
0.002
0.003
0.004
0.005
0.006
0.007
0.2
0.4
0.6
0.8
1.0
alpha for sampling wo rep., alpha / (1-alpha) f

Conclusion

Experiments confirm theoretical results that:

- Bagging always increases squared estimation bias.
- Bagging without replacement with resample size

$$n_{w/o} = \alpha_{w/o} N$$

has the same effect on squared estimation bias and variance as bagging with replacement with resample size

$$n_{with} = \frac{\alpha_{w/o}}{1 - \alpha_{w/o}} N .$$

In fact, agreement is good for individual training samples, not just on average.

Conclusion (continued)

Experiments also support the heuristic that smaller resample size means more smoothing and should lead to smaller variance.

Theory predicts that effect of bagging is $O(1/n^2)$??
Still under investigation.

Thanks for your interest

Conclusion

Experiment confirms theoretical results that:

- Bagging without replacement with resample size

  $$n_{w/o} = \alpha_{w/o} N$$

  has the same effect on squared (estimation) bias, variance, and mean squared (estimation) error as bagging with replacement with resample size

  $$n_{with} = \frac{\alpha_{w/o}}{1 - \alpha_{w/o}} N \, .$$

- Bagging increases squared estimation bias.

In the examples bagging always decreased variance.

Experiment: Bagging regression trees

Same setup as in Friedman and Hall

- $\underline{X} \sim U([0,1]^{10})$
- $Y = f(\underline{X}) + \sigma\epsilon$ with $\epsilon \sim N(0,1)$

Three scenarios:

1. Constant: $f(\underline{x}) = 0$, $\sigma = 1$
2. Piecewise constant: $f(\underline{x}) = \prod_{j=1}^{5} 1(x_j \geq 0.13)$, $\sigma = 0.5$
3. Linear: $f(\underline{x}) = \sum_{j=1}^{5} j\, x_j$, $\sigma = 3$

Training sample sizes $N = 500$ and $N = 5000$

Prediction rule: Cart tree with 50 leaves

Bagging with 50 resamples

Let $p(\underline{x}; \mathcal{X}^{\infty})$ be the rule built from an "infinite" training sample (we use $N = 500,000$)

Quantities of interest

- Variance $\mathbf{E}_{\underline{x}}(\mathbf{var}_{\mathcal{X}}\, p^b(\underline{x}; \mathcal{X}))$
- Squared estimation bias $\mathbf{E}_{\underline{x}}(\mathbf{E}_{\mathcal{X}}\, p^b(\underline{x}; \mathcal{X}) - p(\underline{x}; \mathcal{X}^{\infty}))^2$
- Squared total bias $\mathbf{E}_{\underline{x}}(\mathbf{E}_{\mathcal{X}}\, p^b(\underline{x}; \mathcal{X}) - f(\underline{x}))^2$
- Mean squared error = variance + squared total bias

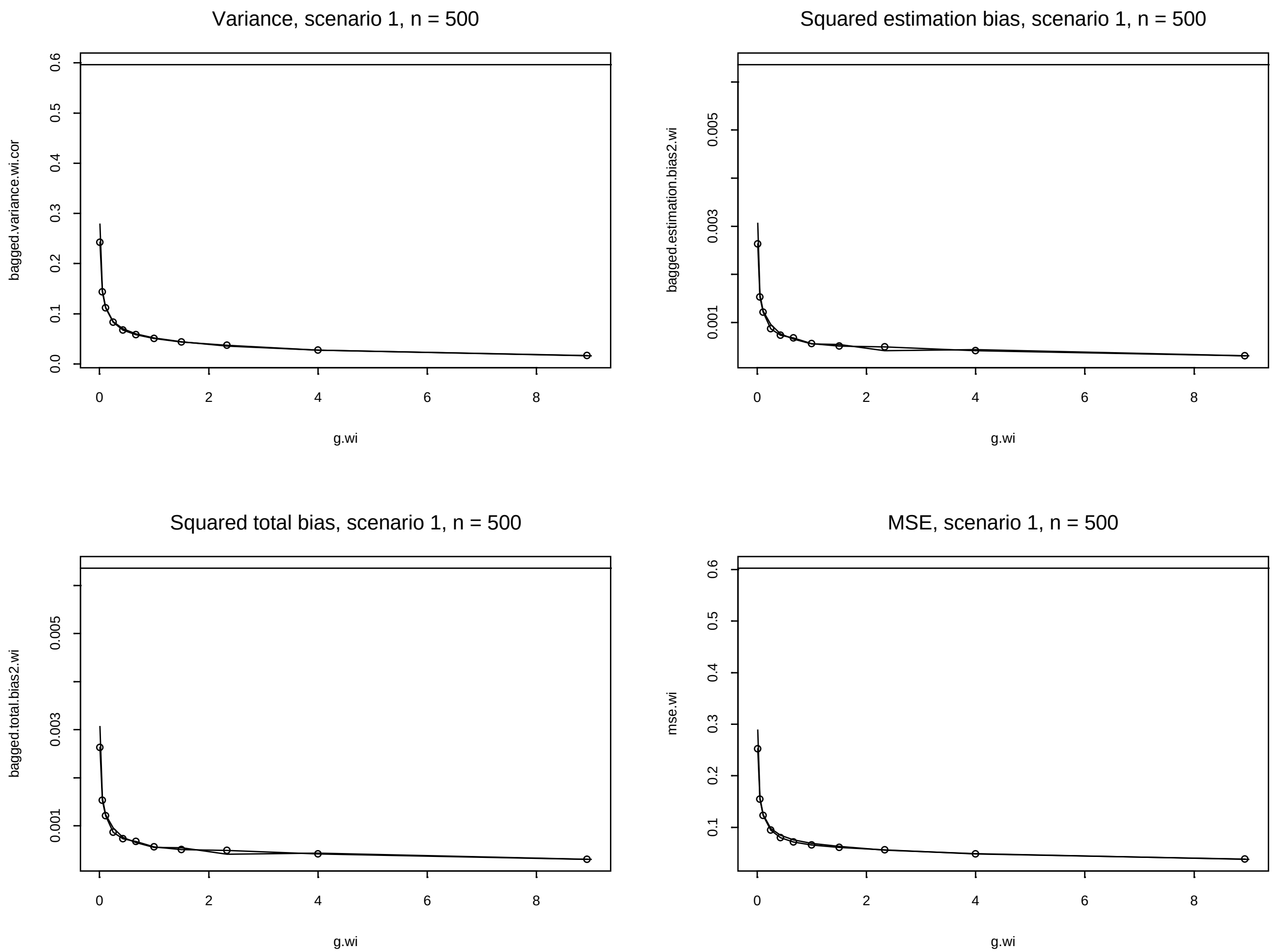
as a function of $g = \frac{N}{n_{with}} = \frac{N}{n_{w/o}} - 1$

Note: Large g means small resample size!

Scenario 1 ($f(\underline{x}) = 0$), $N = 500$

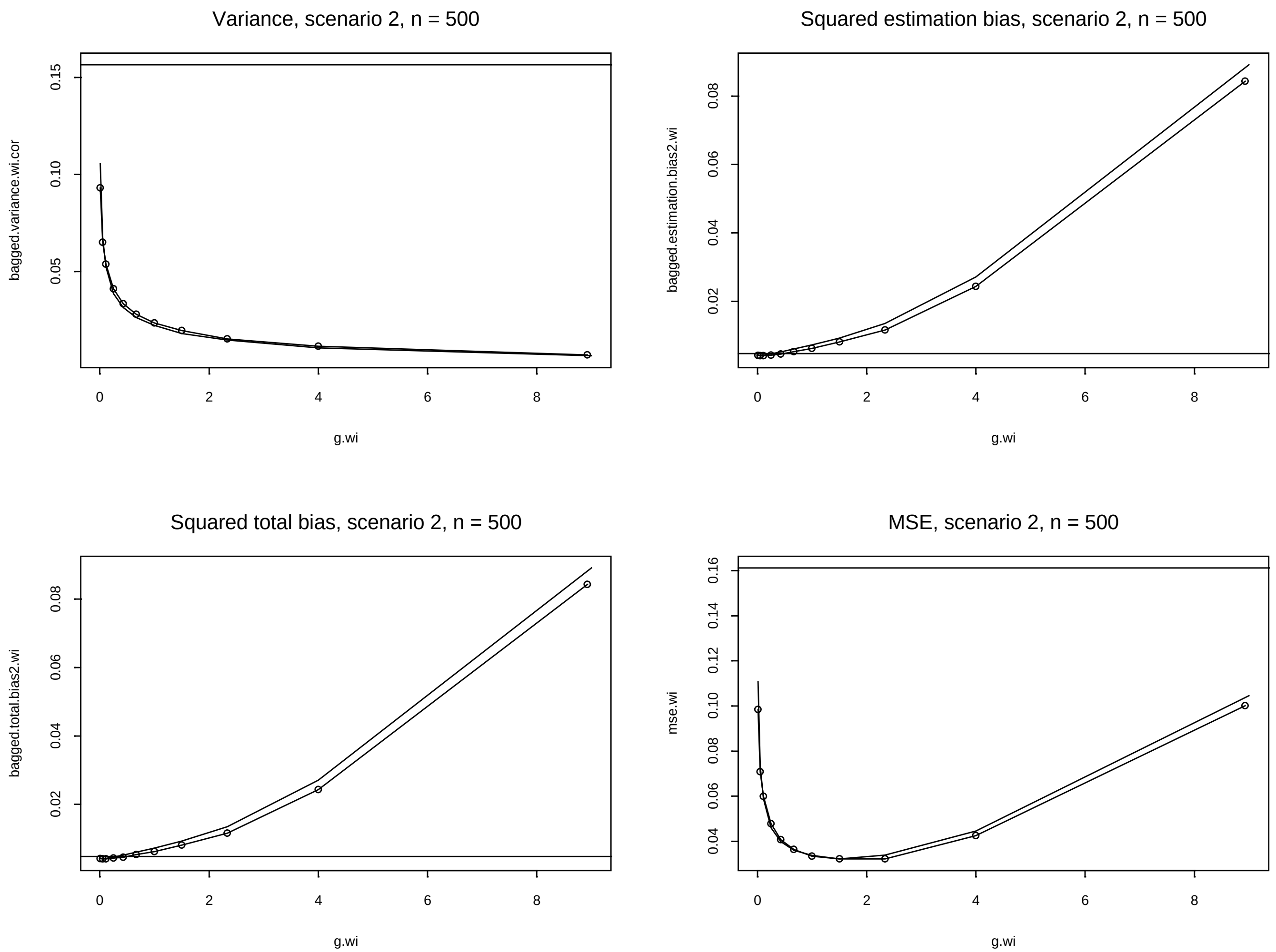
Horizontal lines correspond to unbagged rule.

Note: There are two curves in each plot, for resampling with and without replacement



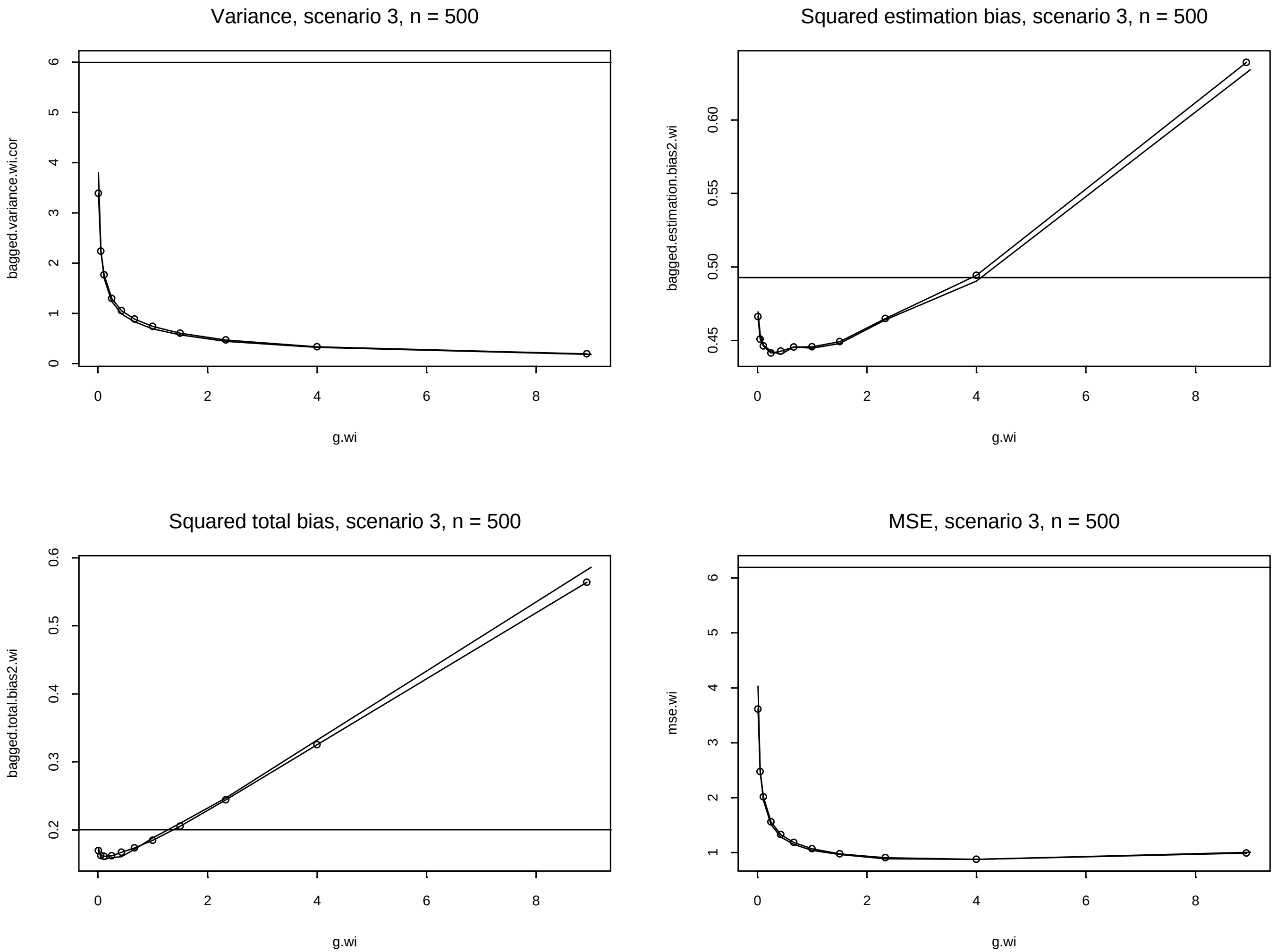
Scenario 2 ($f(\underline{x})$ piecewise constant), $N = 500$

Horizontal lines correspond to unbagged rule.



Scenario 3 ($f(\underline{x})$ linear), $N = 500$

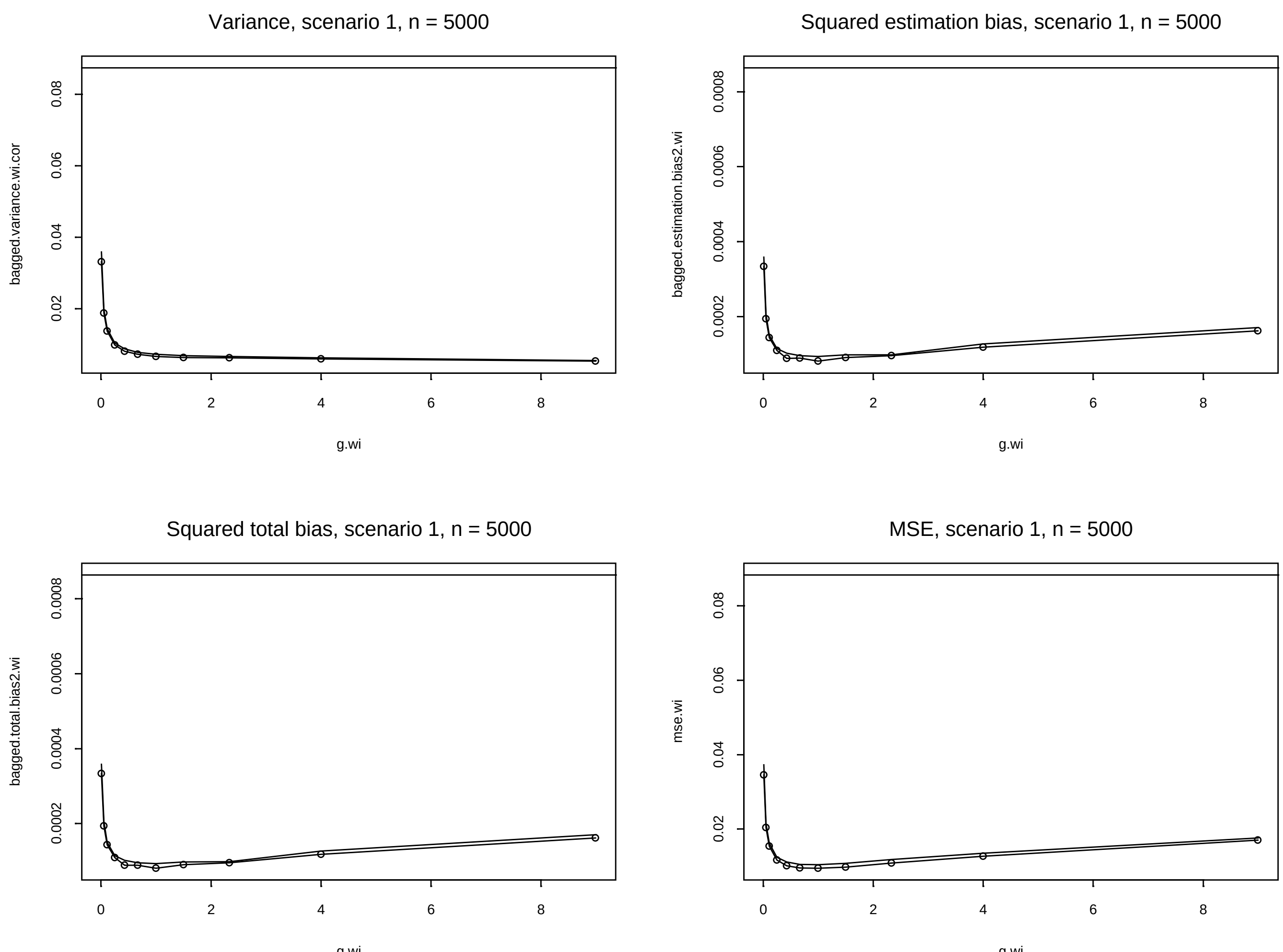
Horizontal lines correspond to unbagged rule.



Scenario 1 ($f(\underline{x}) = 0$), $N = 5000$

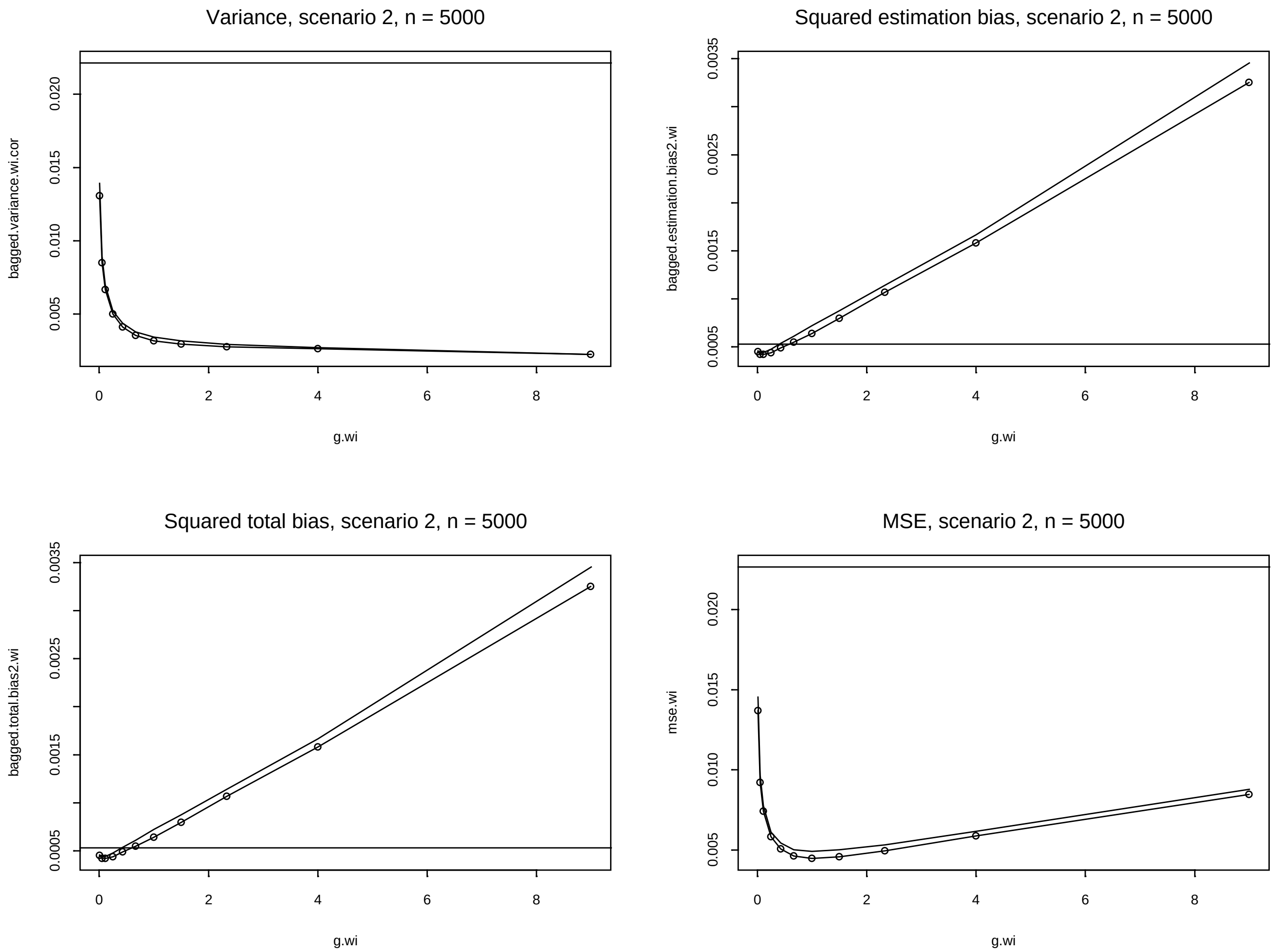
Horizontal lines correspond to unbagged rule.

Comment on increase in MSE



Scenario 2 ($f(\underline{x})$ piecewise constant), $N = 5000$

Horizontal lines correspond to unbagged rule.



Scenario 3 ($f(\underline{x})$ linear), $N = 5000$

Horizontal lines correspond to unbagged rule.