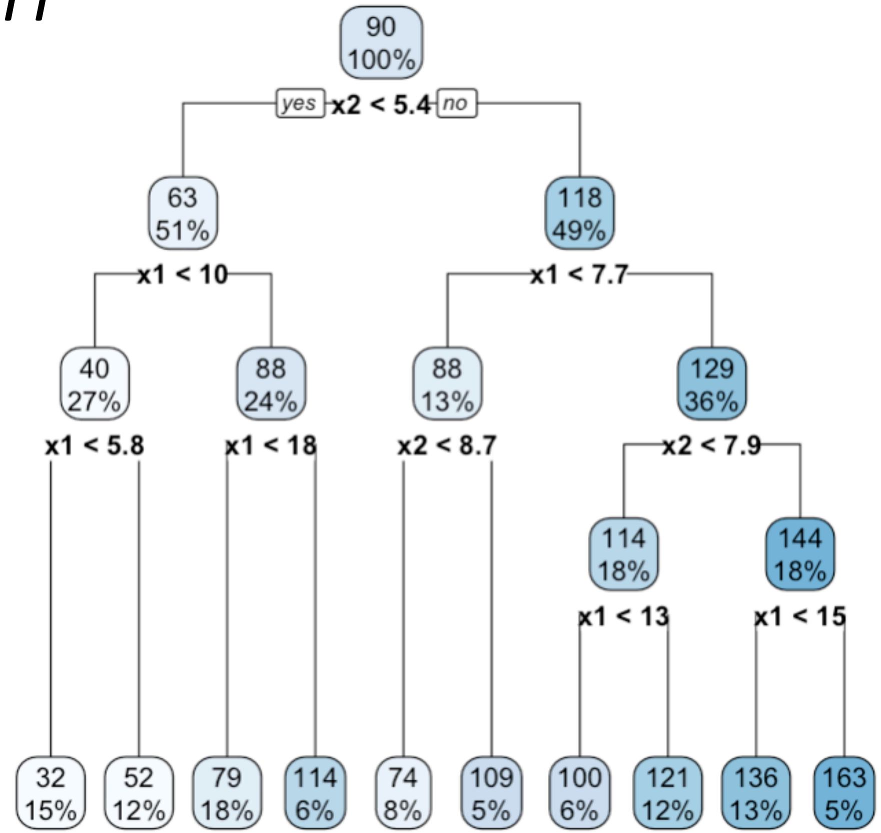
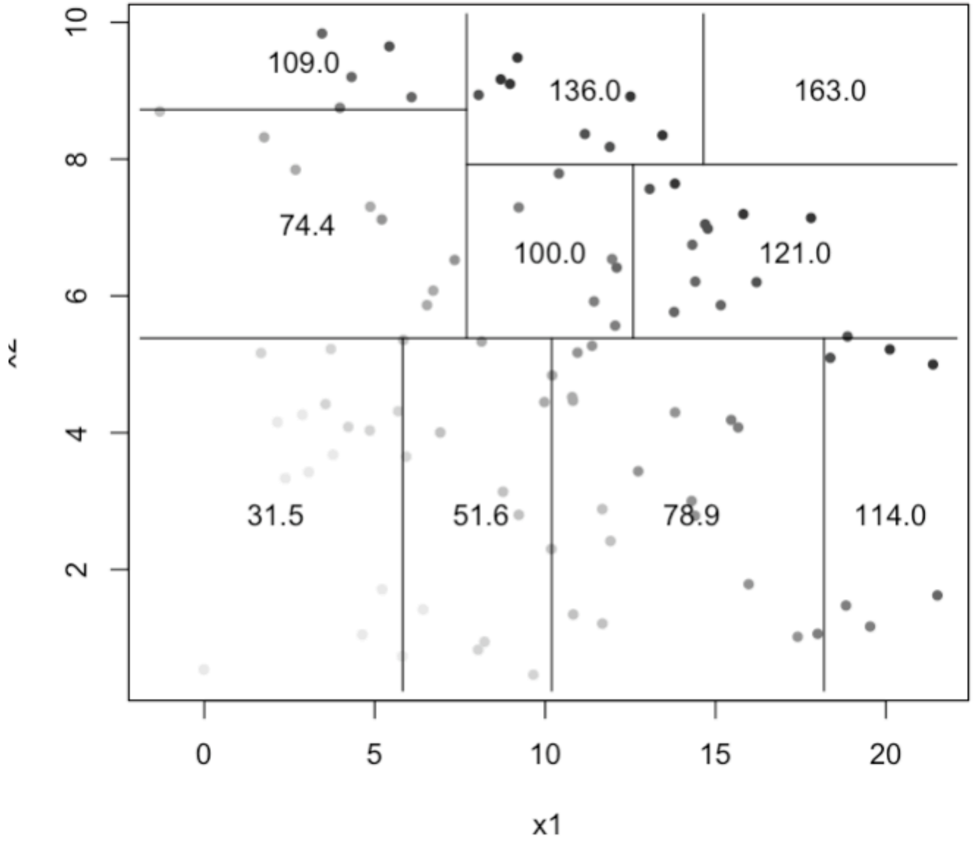
BIOST 527 Final Project: Regression Tree

Yichen Lu, June 2020

Intuition

$$n = 10$$
$$X_1 \sim Normal(10, 5)$$
$$X_2 \sim Uniform(0, 10)$$
$$y = 5x_1 + x_2^2 + \epsilon$$
$$\epsilon \sim Normal(0, 1)$$





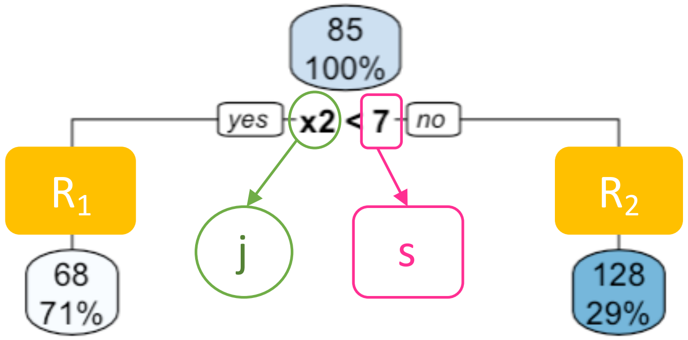
Keywords

- n observations $(x_i, y_i), x_i = (x_{i1}, x_{i2}, \dots x_{ip})$
- M regions $R_1, R_2, \dots R_M$ $N_m = \sum_{i=1}^{n} I(x_i \in R_m)$
- Constant c_m in each region $\hat{f}(x_i) = \sum_{m=1}^{M} c_m I(x_i \in R_m)$
- Minimize the RSS

$$RSS = \sum_{i=1}^{n} (y_i - \hat{f}(x_i))^2 \quad \rightarrow \quad RSS = \sum_{m=1}^{M} \sum_{x_i \in R_m} (y_i - \hat{y}_{R_m})^2 \quad \hat{y}_{R_m} = \frac{1}{N_m} \sum_{x_i \in R_m} y_i$$

- Binary

$$\min_{j,s} \left(\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \right)$$



- Greedy

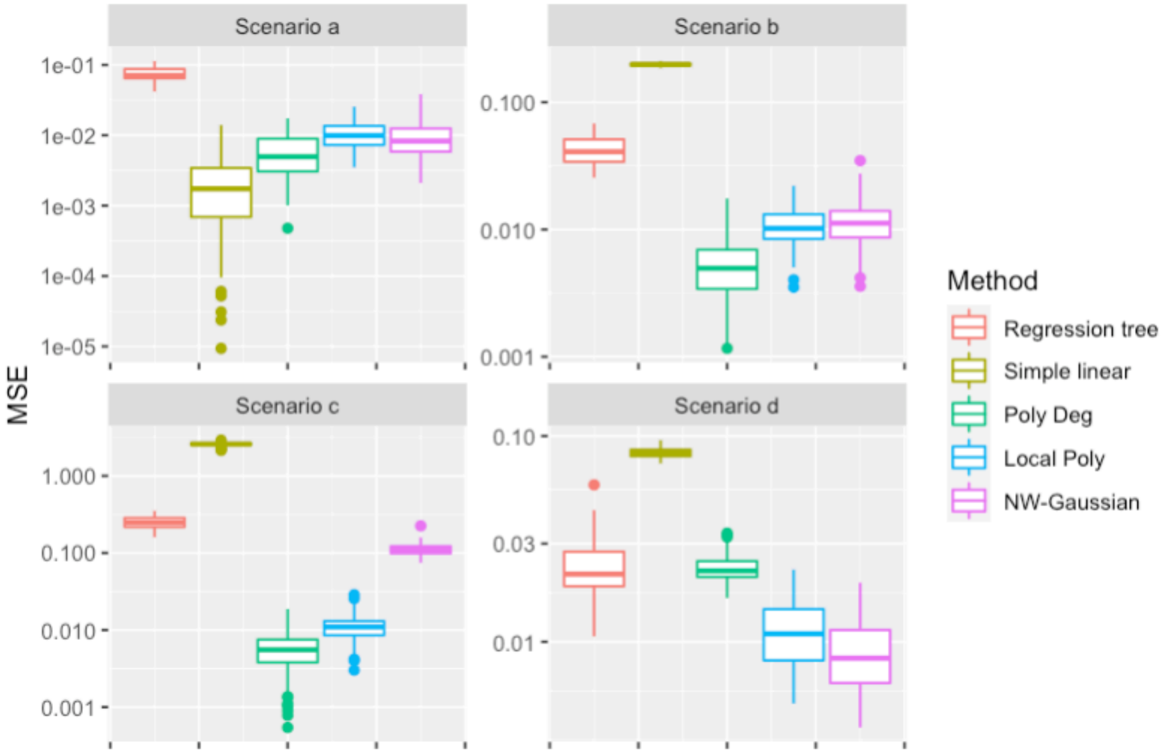
Simulation: Tree

(a) $f(x) = 2x$.

(b) $f(x) = sin(x * \pi)$.

(c) $f(x) = 2x + x^3 - 6x^4$.

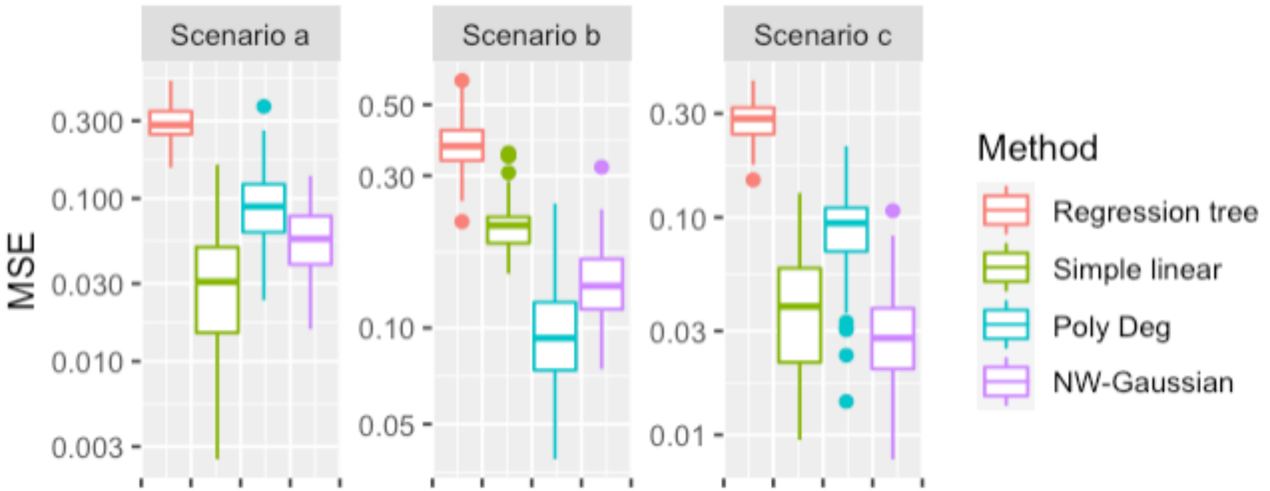
(d) $f(x) = \frac{1}{1+(5x)^2}$.



(a) $f(x_1, x_2, x_3) = x_1 + x_2 + x_3$

(b) $f(x_1, x_2, x_3) = sin(4x_1) + 2\sqrt{x_2} + e^{x_3}$

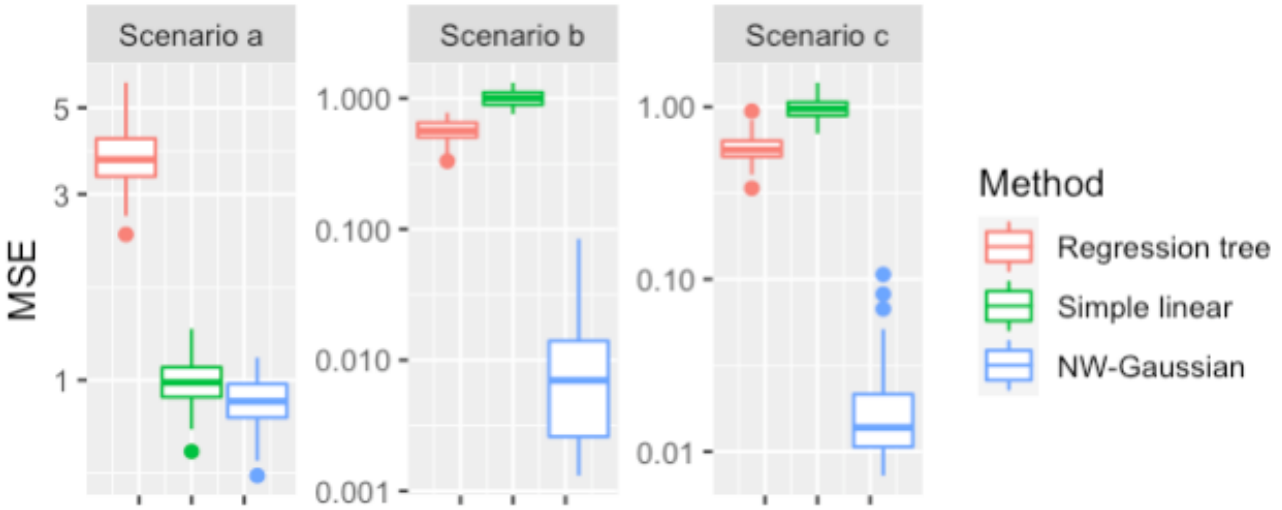
(c) $f(x_1, x_2, x_3) = (x_1 x_2 x_3)^{1/3}$



(a) $f(x) = x_1 + x_2 + \ldots x_{100}$

(b) $f(x) = (x_1 x_2 \ldots x_{100})^{\frac{1}{100}}$

(c) $f(x) \sim Normal(0.5, 0.1)$



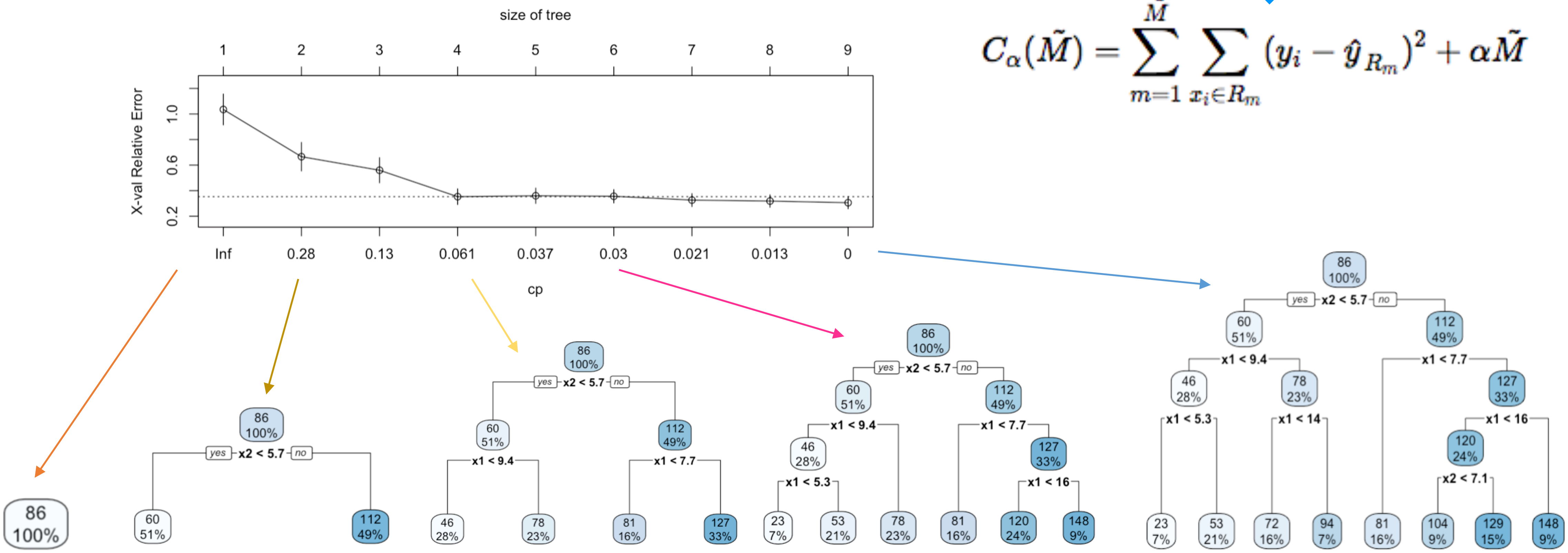
Pruning

- Overfitting
- Methods
  - Only split nodes if decrease in RSS > a threshold
  - Stop splitting when node size < a threshold
  - Cost-complexity/ weakest link pruning

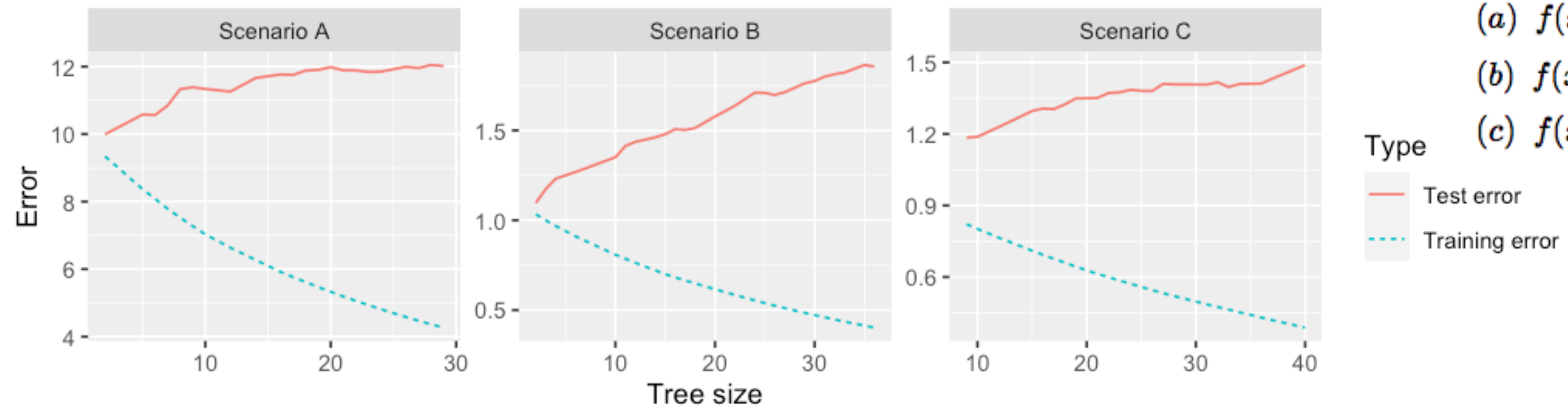
$$RSS = \sum_{m=1}^{M} \sum_{x_i \in R_m} (y_i - \hat{y}_{R_m})^2$$

$$\Big\downarrow$$

$$C_\alpha(\tilde{M}) = \sum_{m=1}^{\tilde{M}} \sum_{x_i \in R_m} (y_i - \hat{y}_{R_m})^2 + \alpha \tilde{M}$$



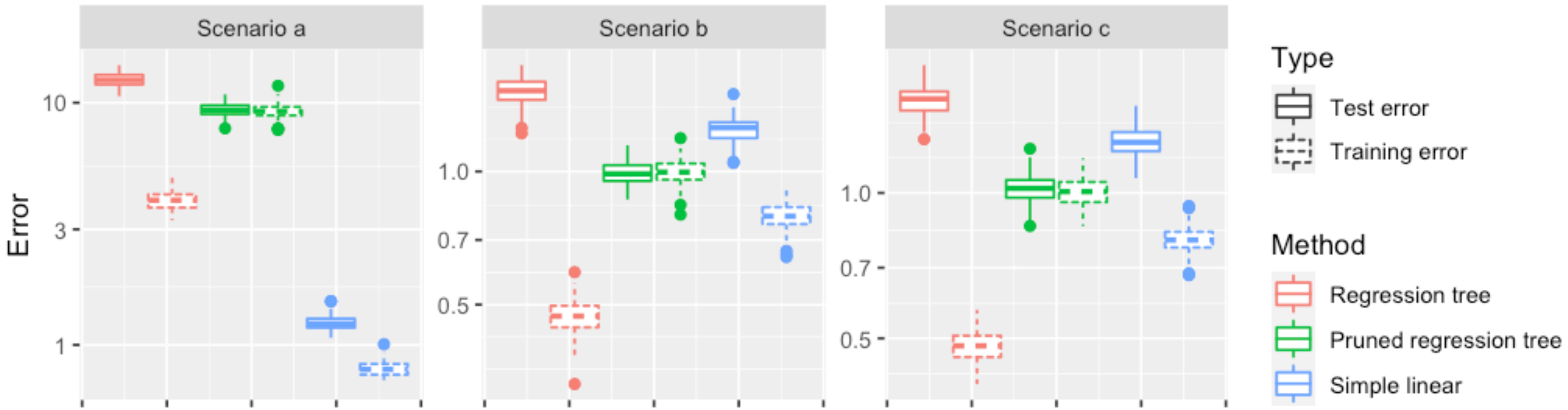
Simulation: Pruning



$$(a)\quad f(x) = x_1 + x_2 + \ldots x_{100}$$

$$(b)\quad f(x) = (x_1 x_2 \ldots x_{100})^{\frac{1}{100}}$$

$$(c)\quad f(x) \sim Normal(0.5, 0.1)$$



References:

[1] James, Gareth, et al. *An introduction to statistical learning*. Vol. 112. New York: springer, 2013.

[2] Friedman, Jerome, Trevor Hastie, and Robert Tibshirani. *The elements of statistical learning*. Vol. 1. No. 10. New York: Springer series in statistics, 2001.

[3] CMU statistics. *Classification and Regression Trees*. 2009.