A comparison of weighted estimators for the population mean

Ye Yang
Weighting in surveys group

Motivation

- Survey sample in which auxiliary variables are known for the population and an outcome variable known only for the sampled and responding units.
- Interest concerns estimating population mean of outcome variable.
- Many methods to estimate the mean, but not sure which to choose from.
- Simulation conducted to assess performance of estimators under different scenarios.

Estimators

- Horvitz-Thompson estimator:

$\hat{\mu}_{HT} = \frac{\sum_{i \in r} w_i y_i}{N}$

- Robust Horvitz-Thompson estimator:

$B^{HT}_i = (w_i - 1)\left\{y_i - \pi_i \frac{\sum_{j \in r} y_j(1-\pi_j)/\pi_j}{\sum_{j \in r}(1-\pi_j)}\right\}, i \in r$

$\hat{\mu}_{RHT} = \hat{\mu}_{HT} - (B^{HT}_{Min} + B^{HT}_{Max}) / (2N)$

- Horvitz-Thompson estimator with post-stratification:

$\hat{\mu}_{HTPS} = \sum_{i \in r} w_i y_i (N_j / \widehat{N_j})$, where i belongs to post-stratum j and $\widehat{N_j} = \sum_{i \in r_j} w_i$

Estimators

- Hajek estimator:

$$\hat{\mu}_{HA} = \frac{\sum_{i \in r} w_i y_i}{\sum_{i \in r} w_i}$$

- Robust Hajek estimator:

$$B^{HA}_i = (w_i - 1)\left\{e_i - \pi_i \frac{\sum_{j \in r} y_j(1-\pi_j)/\pi_j}{\sum_{j \in r}(1-\pi_j)}\right\},\ i \in r,\ e_i = y_i - \hat{\mu}_{\mathrm{HA}}$$

$$\hat{\mu}_{RHA} = \hat{\mu}_{\mathrm{HA}} - (B^{HA}_{\mathrm{Min}} + B^{HA}_{\mathrm{Max}}) / (2N)$$

- Trimmed estimator:

$$\hat{\mu}_{TR} = \frac{\sum_{i \in r} w^*_i y_i}{\sum_{i \in r} w^*_i}$$

$$w^*_i = \begin{cases} w_0 \, if \ w_i \geq w_0 \\ \gamma w_i \ if \ w_i < w_0 \end{cases}$$, where $w_0 = 3\overline{w}$, $\gamma = \frac{N - \sum_{i \in r} k_i w_0}{\sum_{i \in r}(1-k_i) w_i}$, and $k_i = 1\{w_i \geq w_0\}$

Estimators

- Beaumont estimator:

$\hat{\mu}_{BM} = \frac{\sum_{i \in r} \tilde{w}_i y_i}{\sum_{i \in r} \tilde{w}_i}$, where $\tilde{w}_i$ is obtained by regressing w_i on a spline of y_i

- Penalized spline of propensity prediction (PSPP)

$y_i = \beta_0 + \beta_1 \pi_i + \sum_{k=1}^{K} \gamma_k (\pi_i - \kappa_k)_+ + \varepsilon_i$, $\varepsilon_i \sim N(0, \pi_i^{2\alpha} \sigma^2)$

$(\pi_i - \kappa_k)_+ = \begin{cases} (\pi_i - \kappa_k) \text{ if } (\pi_i - \kappa_k) > 0 \\ 0 \text{ otherwise} \end{cases}$

$\hat{\mu}_{PSPP} = \frac{\sum_{i \in U} \hat{y}_i}{N}$

Questions to answer

- How does diving the weighted sum by N as in the Horvitz-Thompson estimator compare with dividing by $\sum_{i \in r} w_i$ as in the Hajek estimator?
- How do the robust versions of HT and Hajek improve upon the respective estimators?
- How does weight trimming and the Beaumont method improve upon the Hajek estimate?
- How do the weighted estimators compare with and PSPP?
- Impact of variance structure on PSPP.

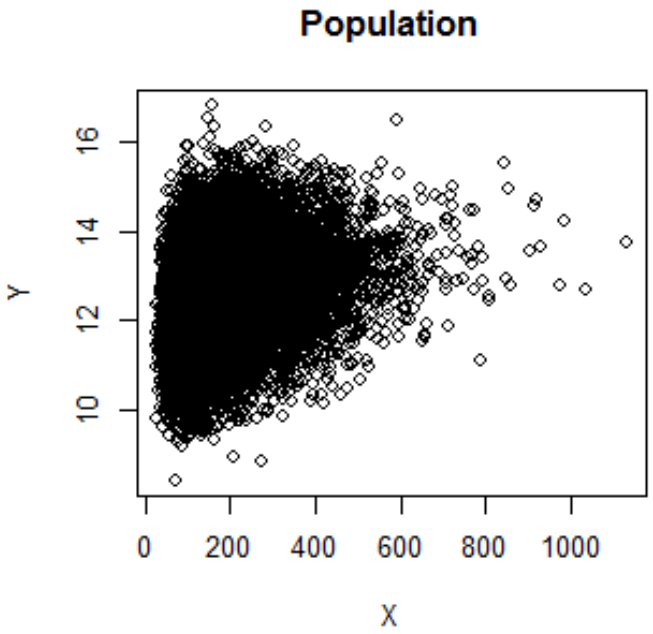
Simulations

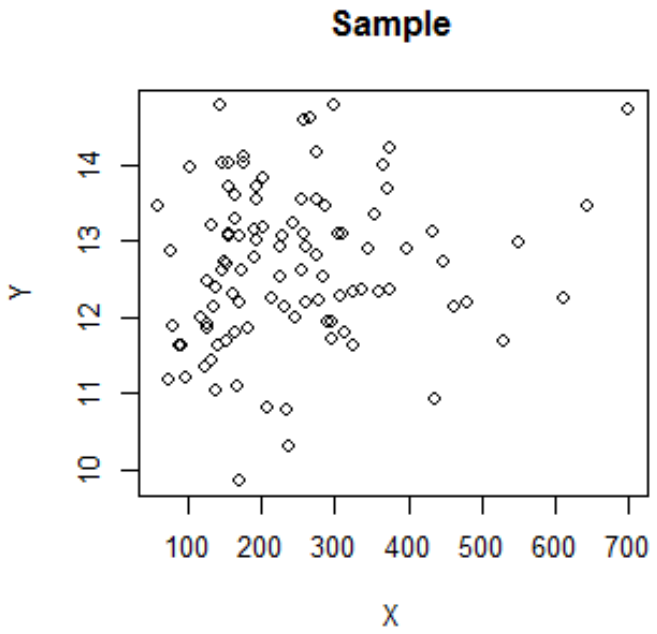
- 10 scenarios covering stratified PPS sampling, SRS with nonresponse, and PPS sample with nonresponse.
- 500 replications.
- Performance of estimators compared using relative root mean squared error (RRMSE):

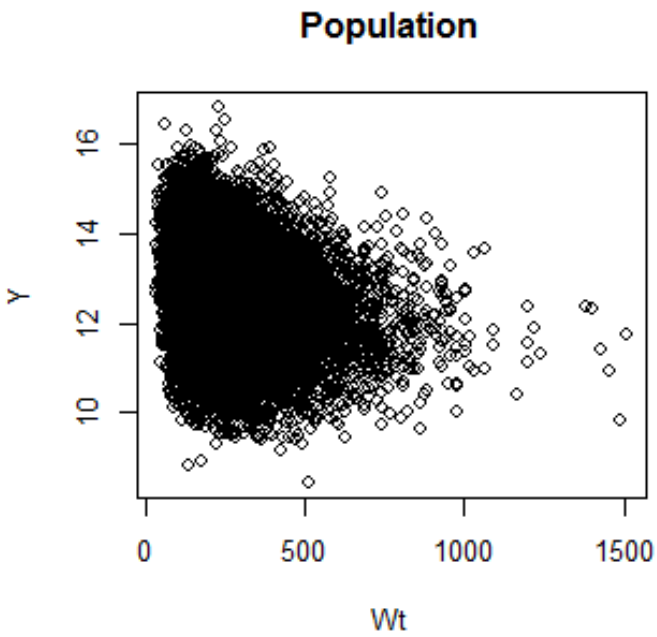
$$RRMSE_{estimator} = 100 * \frac{RMSE_{estimator}}{RMSE_{Hajek}}$$

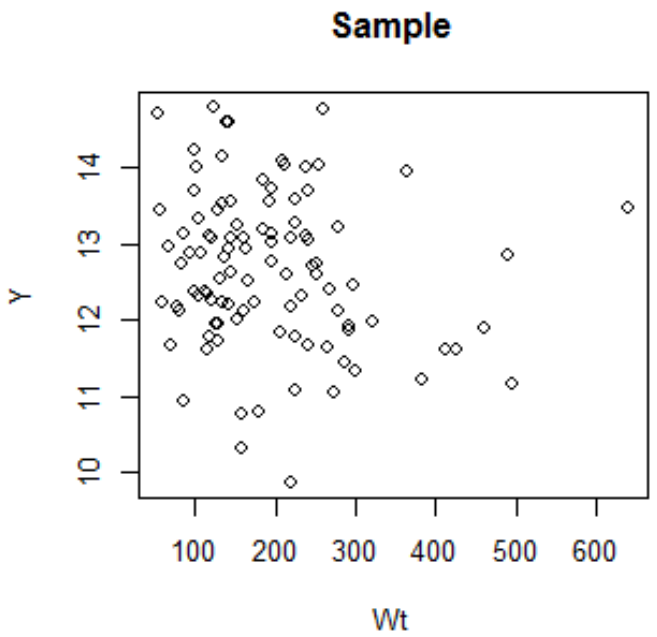
Scenario 1

- Population:
  - Three strata: Z = 1, 2, 3
  - Strata population sizes N_1=5,000, N_2=6,000, N_3=9,000
  - $X|X_1$ = floor($100X_1$), $\log(X_1) \sim N(0.5, 0.5)$
- Sample:
  - n = 100, 500
  - Proportional allocation of strata
  - Stratified PPS sampling with X as size variable
- Weights:
  - Selection probability $\pi_{hi} = n_h X_{hi} / \sum_{i=1}^{N_h} X_{hi}$
  - Sample weight $w_{hi} = 1 / \pi_{hi}$
- $Y|X \sim N(10 + 0.5\log X, 1)$

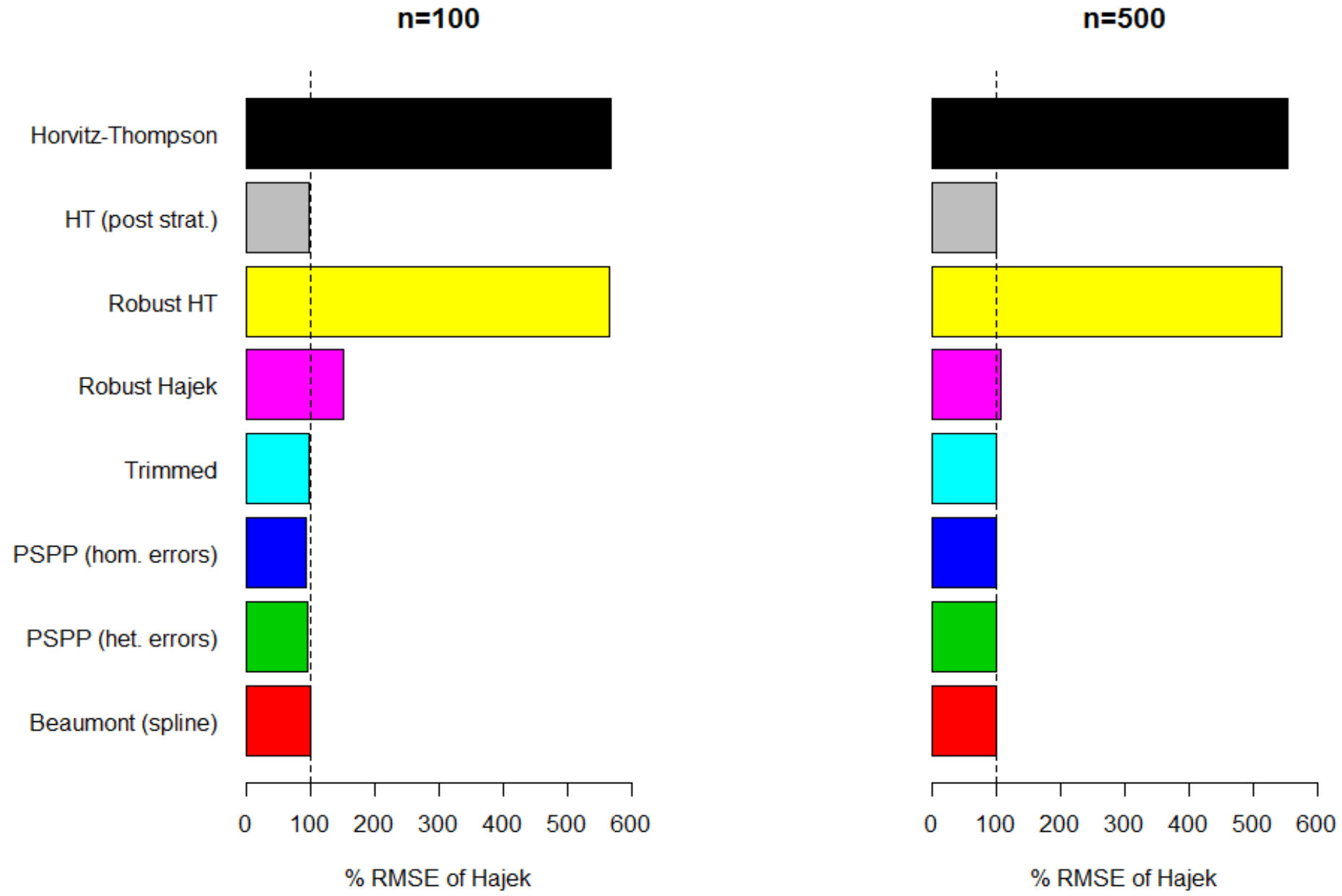

Population
Y
X
10
12
14
16
0
200
400
600
800
1000


Sample
Y
X
10
11
12
13
14
100
200
300
400
500
600
700


Population
Y
Wt
10
12
14
16
0
500
1000
1500

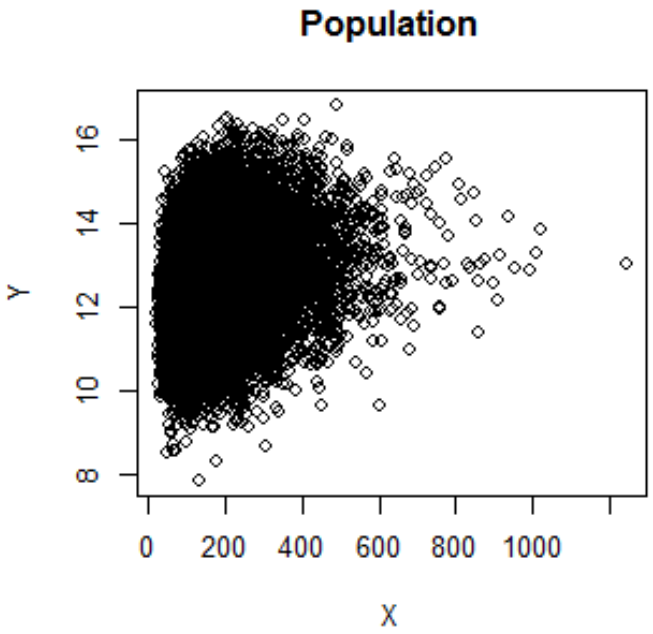

Sample
Y
Wt
10
11
12
13
14
100
200
300
400
500
600

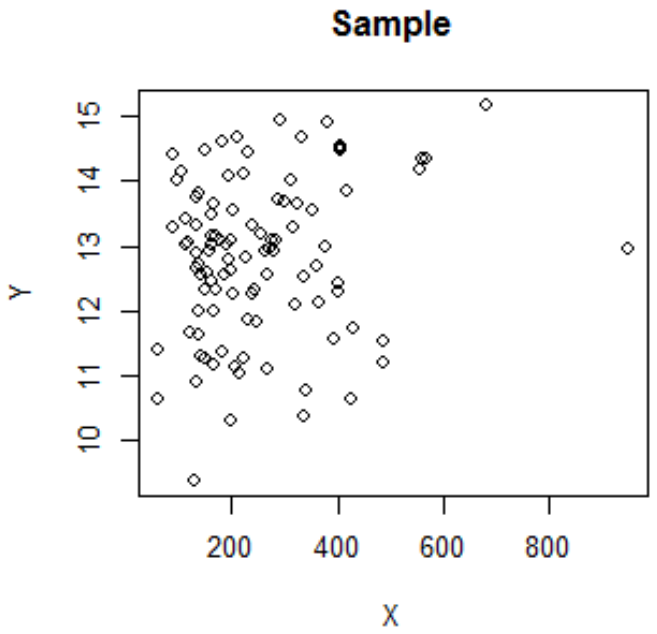
Relative RMSE for Scenario 1

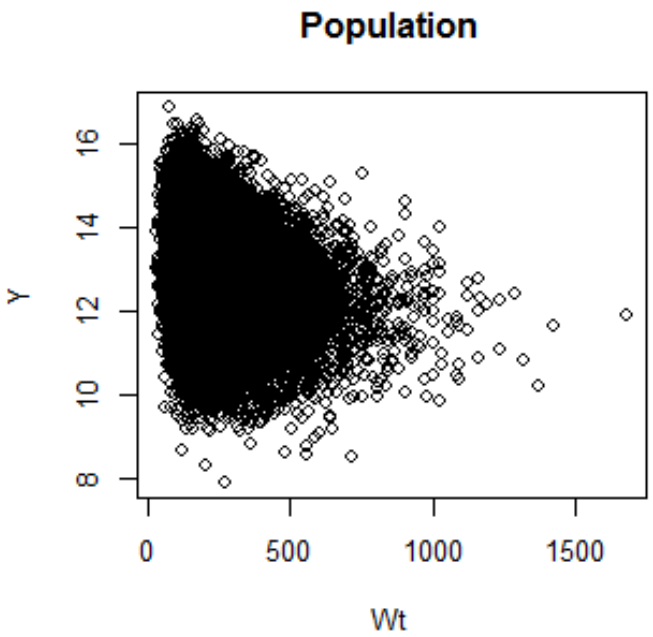


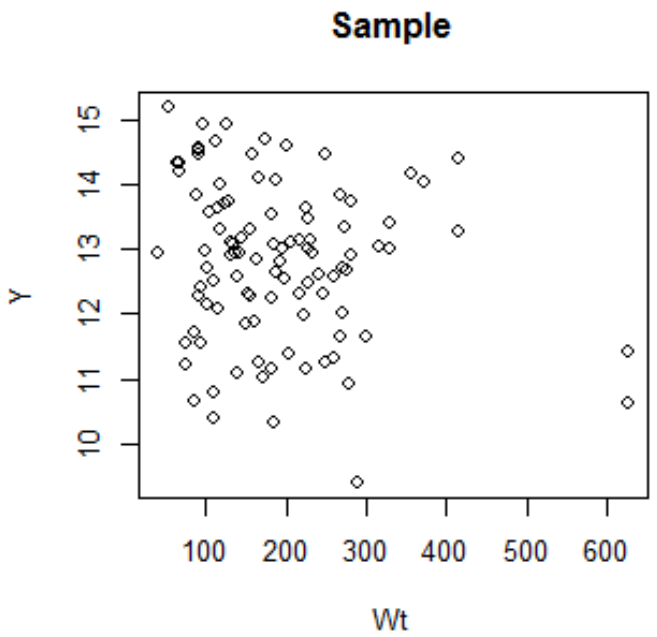
Scenario 2

- Population:
  - Three strata: Z = 1, 2, 3
  - Strata population sizes N_1=5,000, N_2=6,000, N_3=9,000
  - $X|X_1$ = floor($100X_1$), $\log(X_1) \sim N(0.5, 0.5)$
- Sample:
  - n = 100, 500
  - Proportional allocation of strata
  - Stratified PPS sampling with X as size variable
- Weights:
  - Selection probability $\pi_{hi} = n_h X_{hi} / \sum_{i=1}^{N_h} X_{hi}$
  - Sample weight $w_{hi} = 1 / \pi_{hi}$
- $Y|Z, X \sim N(10 + 0.5\log X + 0.5\{Z=2\} - \{Z=3\} - 0.2\{Z=2\}\log X + 0.3\{Z=3\}\log X, 1)$

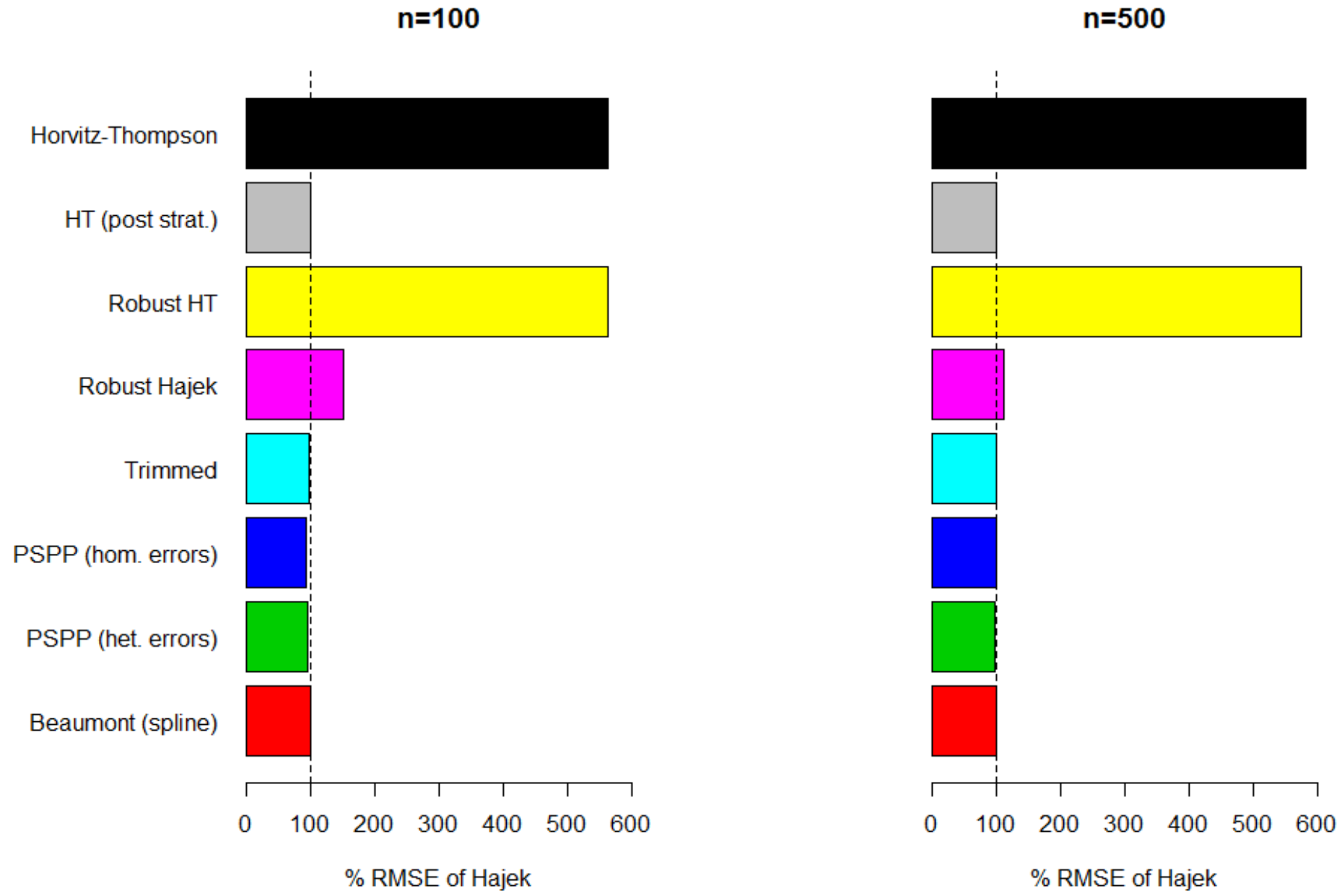

Population
Y
X
8
10
12
14
16
0
200
400
600
800
1000


Sample
Y
X
10
11
12
13
14
15
200
400
600
800


Population
Y
Wt
8
10
12
14
16
0
500
1000
1500

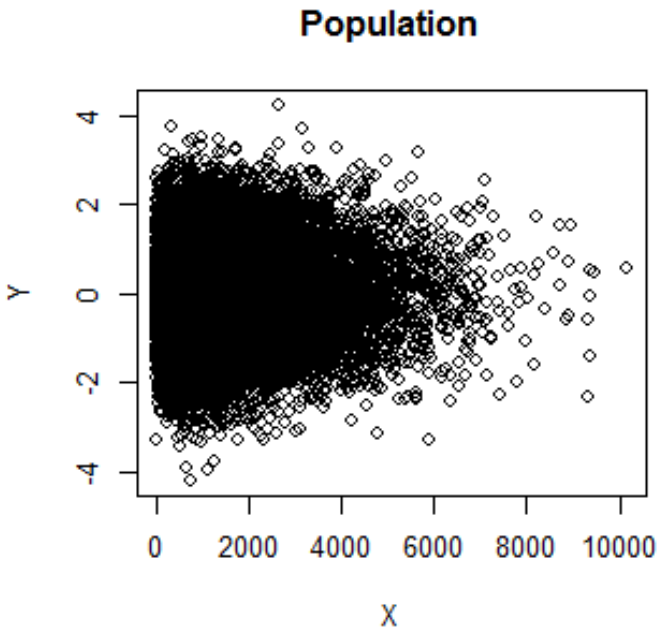

Sample
Y
Wt
10
11
12
13
14
15
100
200
300
400
500
600

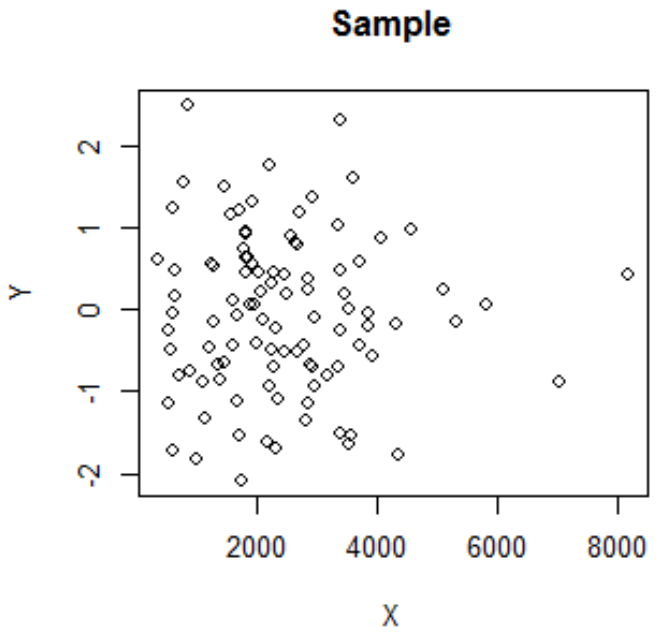
Relative RMSE for Scenario 2

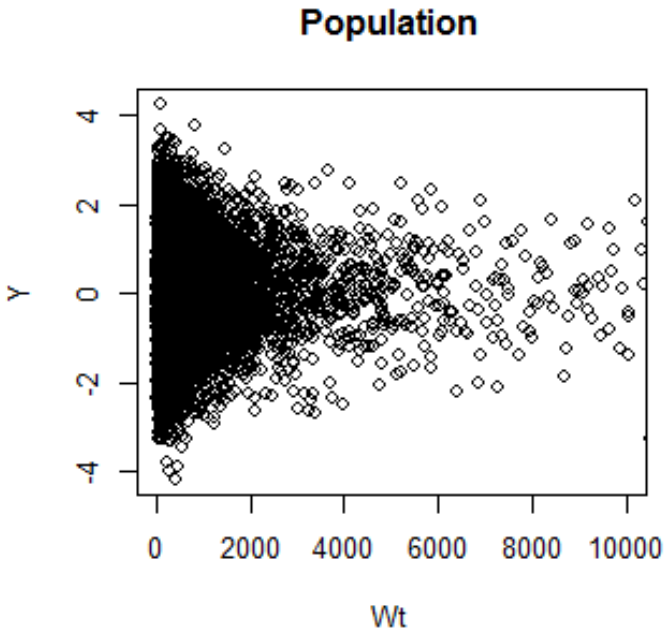


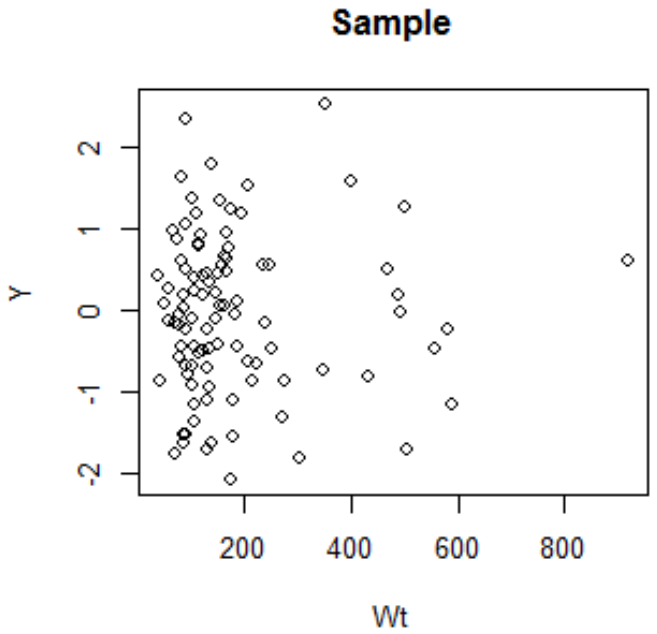
Scenario 3

- Population:
  - N = 20,000
  - X ~ GAMMA(1.5, 0.001)
- Sample:
  - n = 100, 500
  - PPS sampling with X as size variable
- Weights:
  - Selection probability $\pi_i = nX_i / \sum_{i=1}^{N} X_i$
  - Sample weight $w_i = 1 / \pi_i$
- Y|X ~ N(0, 1)

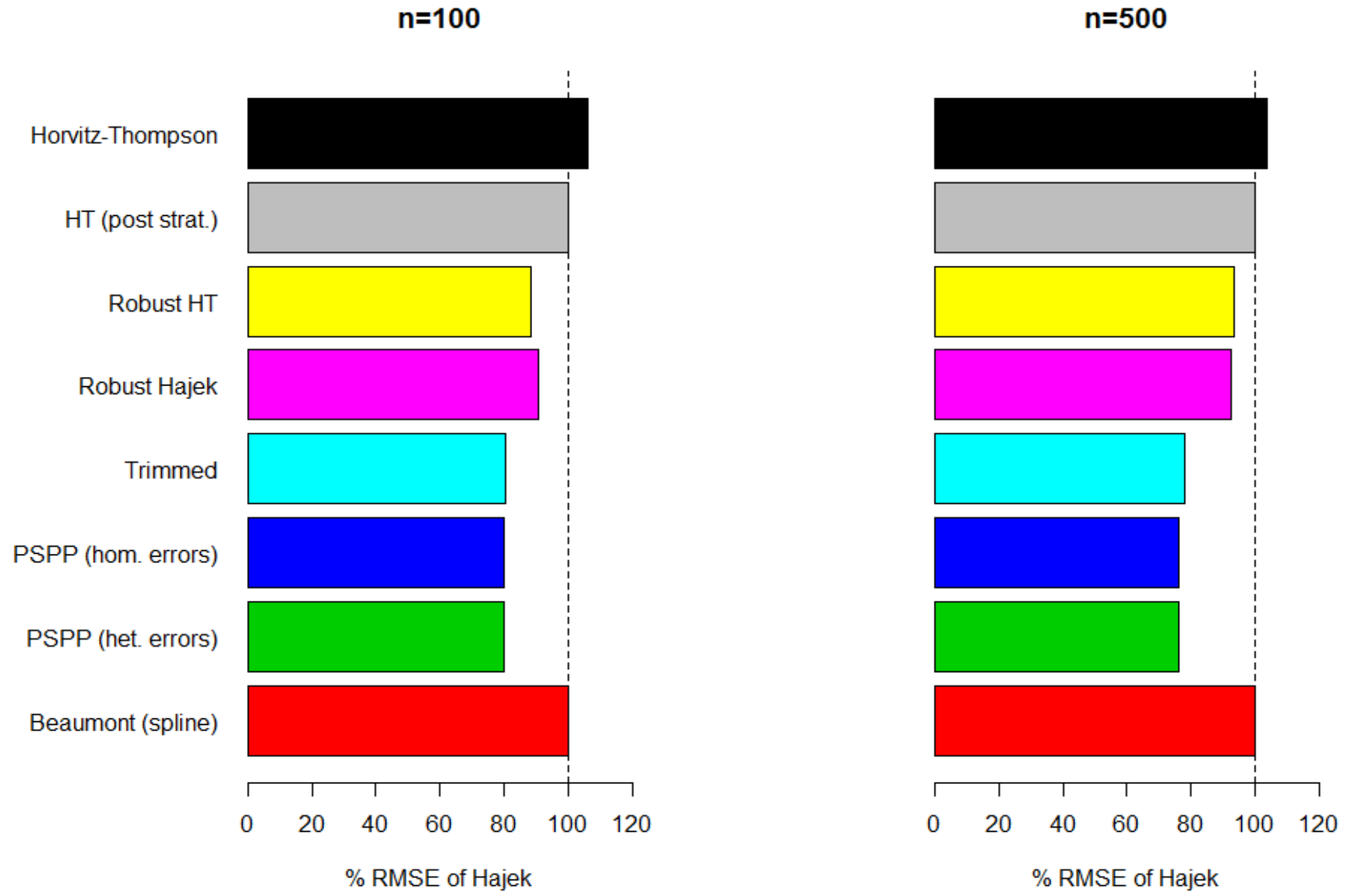

Population
Y
X
4
2
0
-2
-4
0
2000
4000
6000
8000
10000


Sample
Y
X
2
1
0
-1
-2
2000
4000
6000
8000


Population
Y
Wt
4
2
0
-2
-4
0
2000
4000
6000
8000
10000

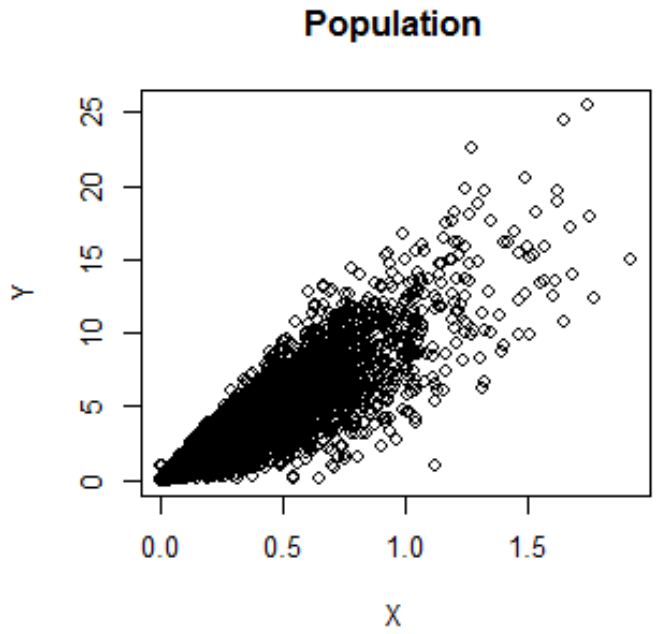

Sample
Y
Wt
2
1
0
-1
-2
200
400
600
800

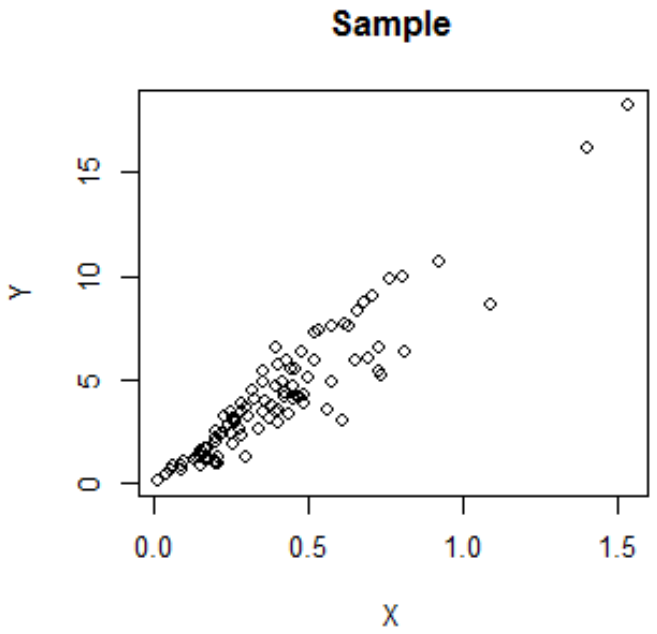
Relative RMSE for Scenario 3

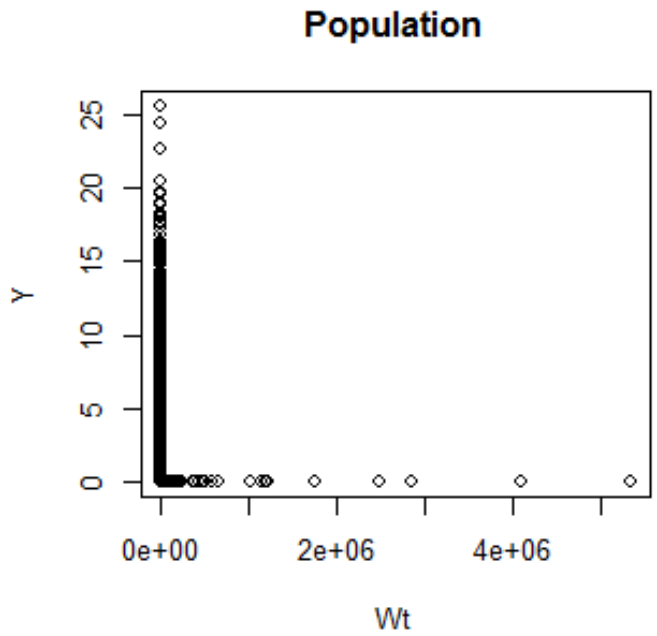


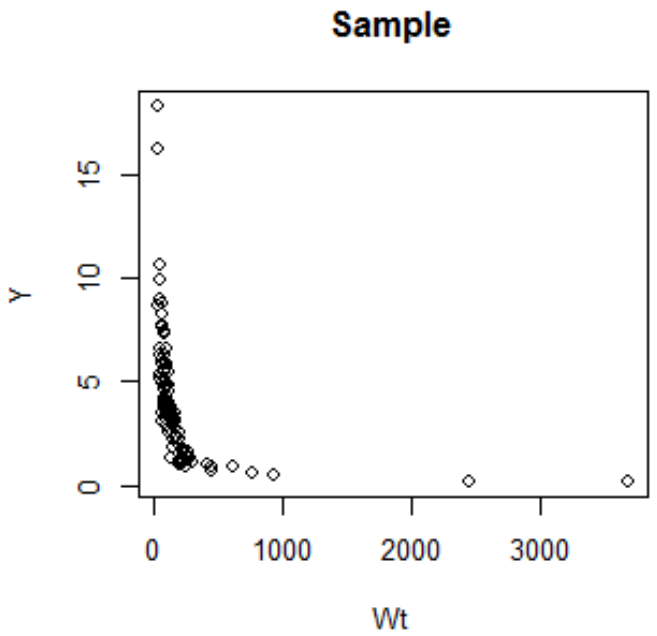
Scenario 4

- Population:
  - N = 20,000
  - X $\sim$ GAMMA(1.5, 0.001)
  - e $\sim$ N(0, 1)
  - Y|X, e = 10*X + 3*X*e
  - If Y $\leq$ 0 then Y = 1
- Sample:
  - n = 100, 500
  - PPS sampling with X as size variable
- Weights:
  - Selection probability $\pi_i = nX_i / \sum_{i=1}^{N} X_i$
  - Sample weight $w_i = 1 / \pi_i$

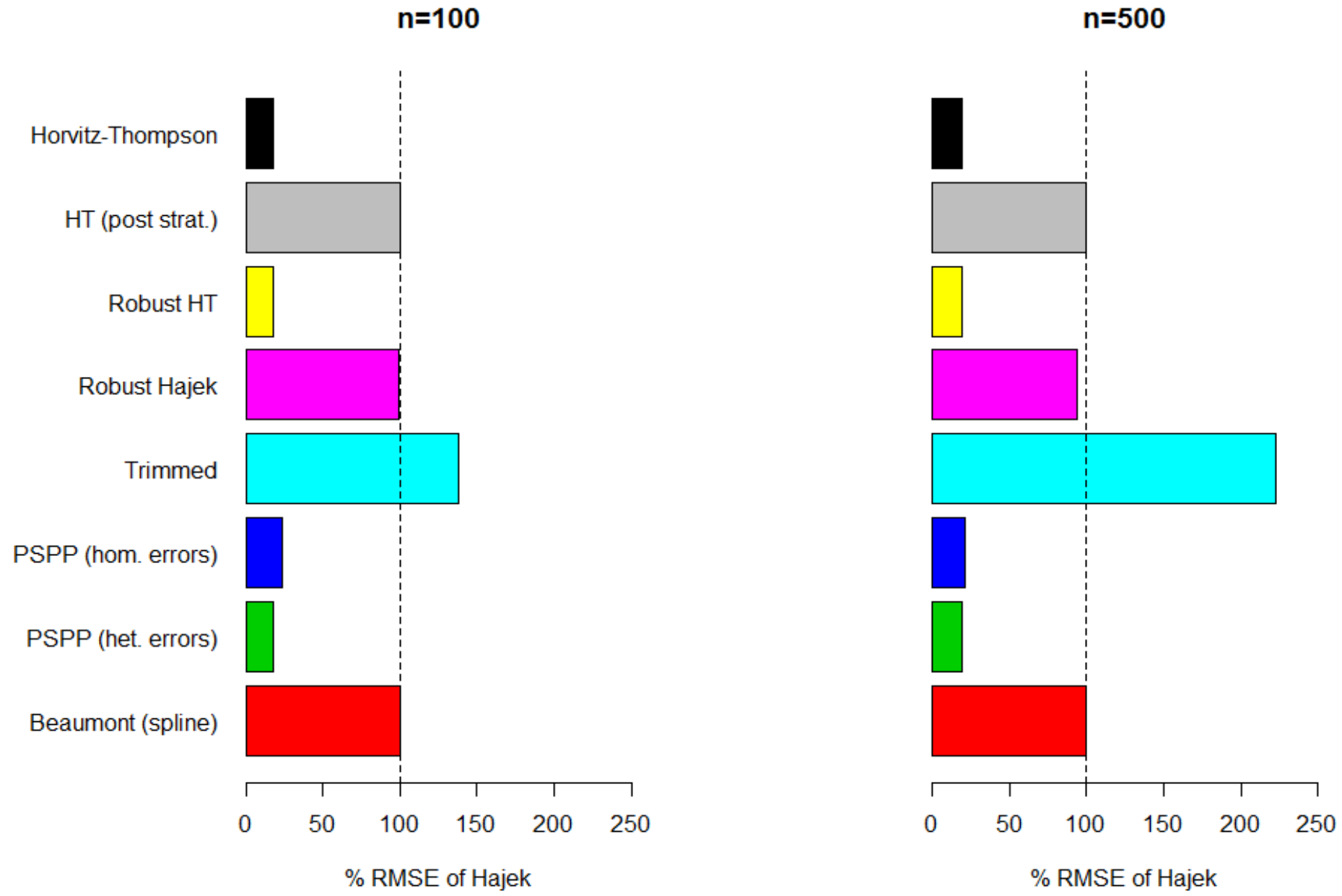

Population
Y
X
0
5
10
15
20
25
0.0
0.5
1.0
1.5


Sample
Y
X
0
5
10
15
0.0
0.5
1.0
1.5


Population
Y
Wt
0
5
10
15
20
25
0e+00
2e+06
4e+06

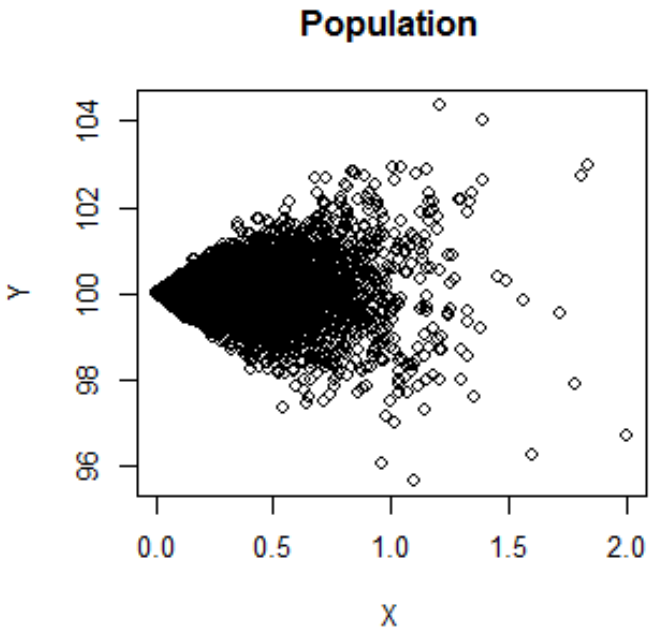

Sample
Y
Wt
0
5
10
15
0
1000
2000
3000

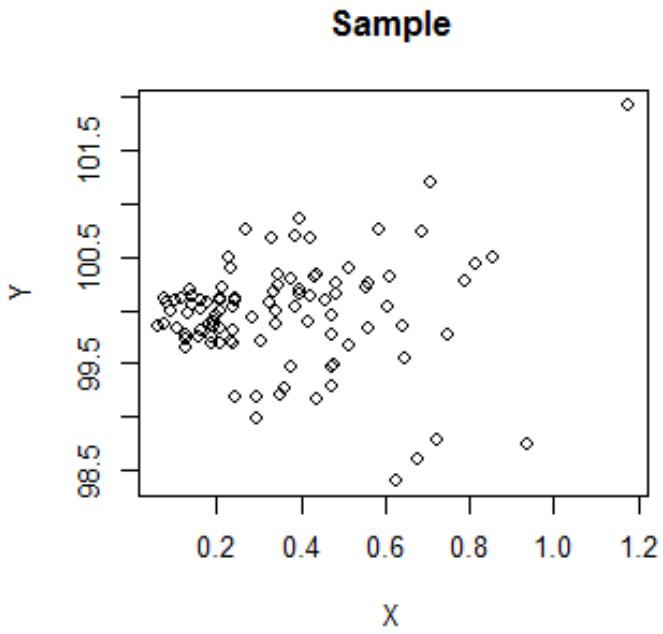
Relative RMSE for Scenario 4

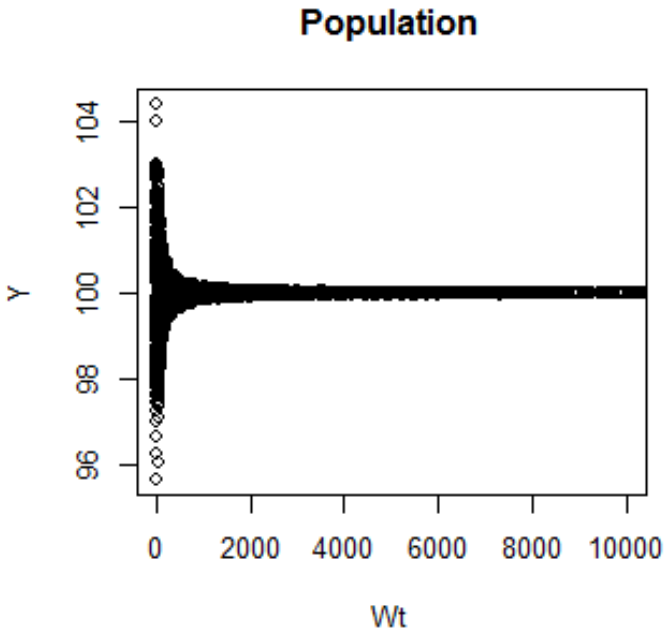


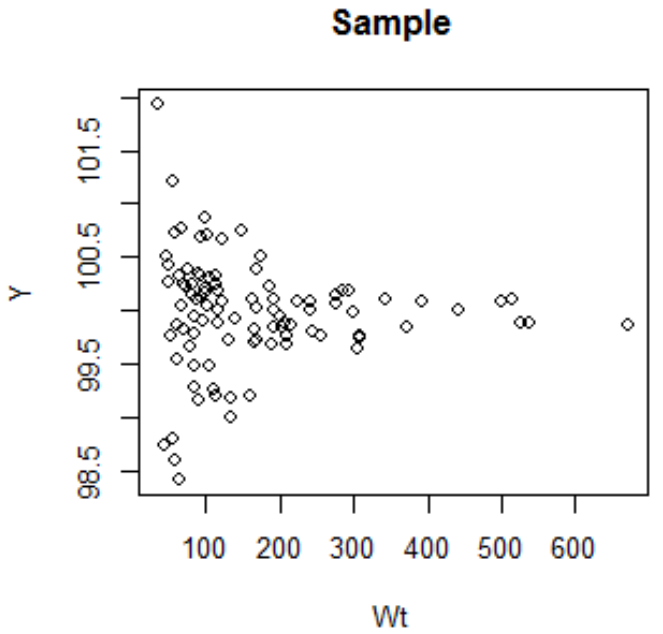
Scenario 5

- Population:
  - N = 20,000
  - X ~ GAMMA(1.5, 0.001)
  - e ~ N(0, 1)
  - Y|X, e = 100 + 3*X*e
- Sample:
  - n = 100, 500
  - PPS sampling with X as size variable
- Weights:
  - Selection probability $\pi_i = nX_i / \sum_{i=1}^{N} X_i$
  - Sample weight $w_i = 1 / \pi_i$

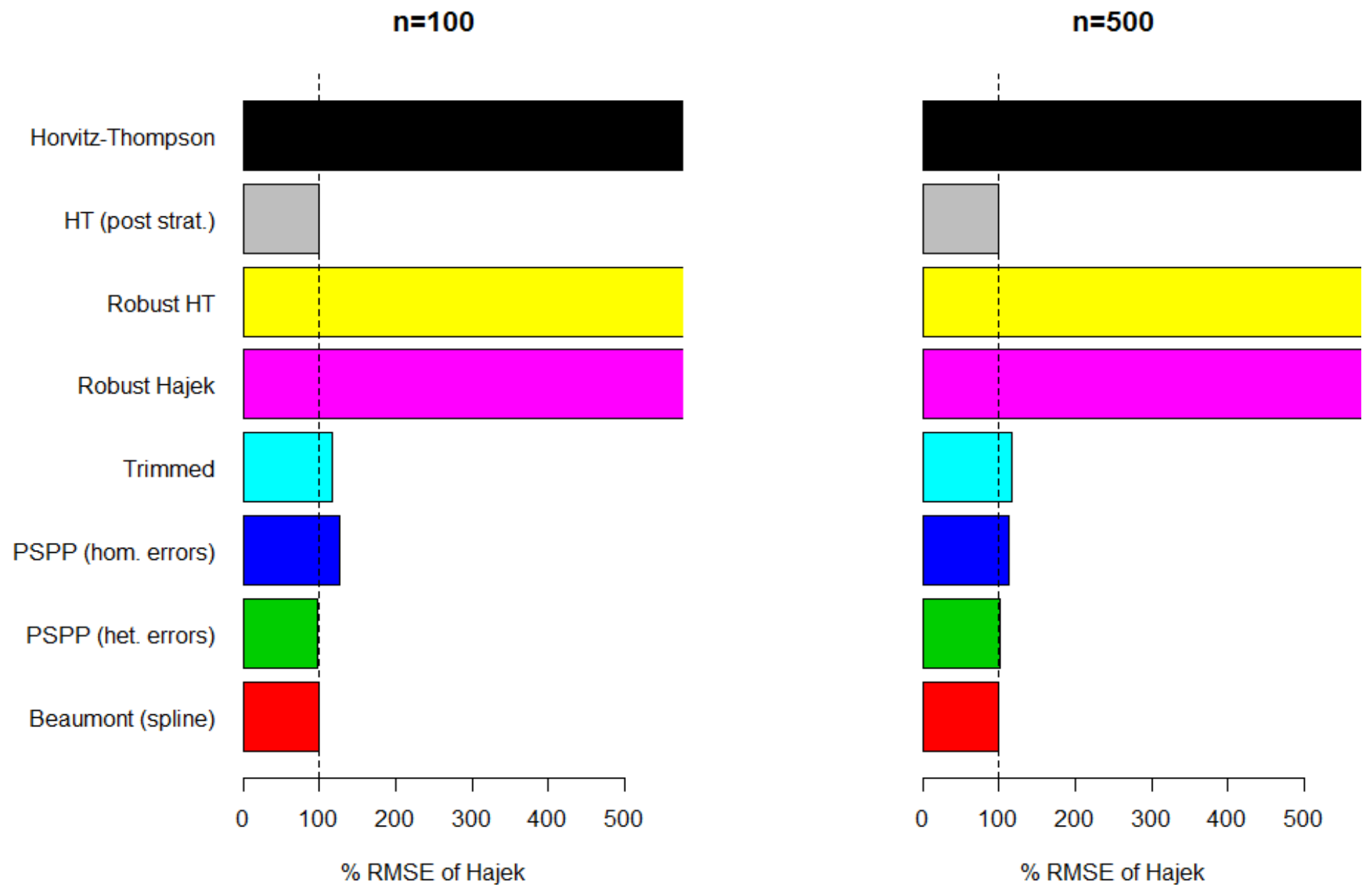

Population
Y
X
96
98
100
102
104
0.0
0.5
1.0
1.5
2.0


Sample
Y
X
98.5
99.5
100.5
101.5
0.2
0.4
0.6
0.8
1.0
1.2


Population
Y
Wt
96
98
100
102
104
0
2000
4000
6000
8000
10000

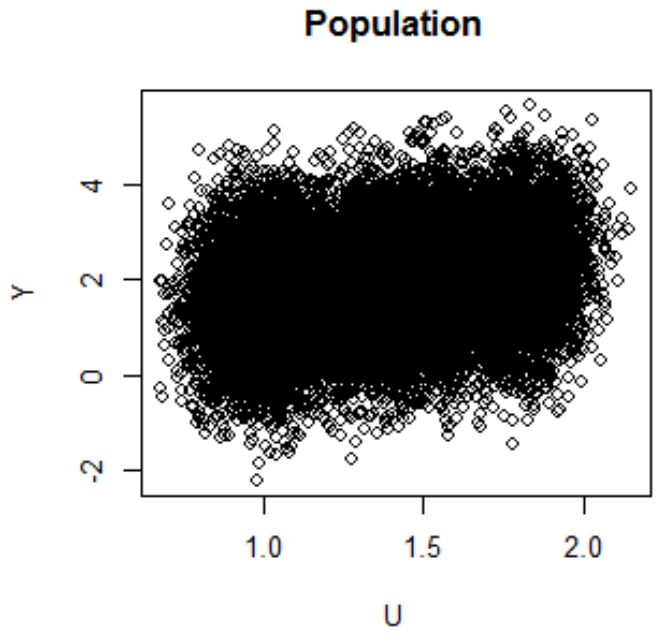

Sample
Y
Wt
98.5
99.5
100.5
101.5
100
200
300
400
500
600

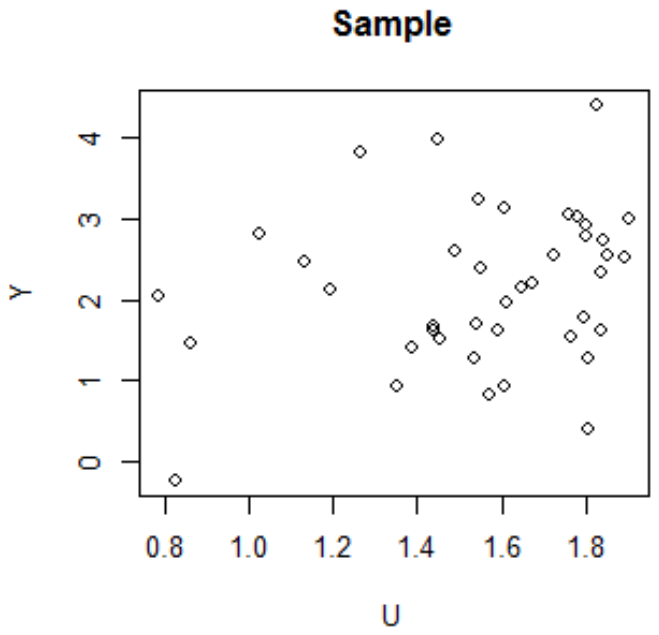
Relative RMSE for Scenario 5

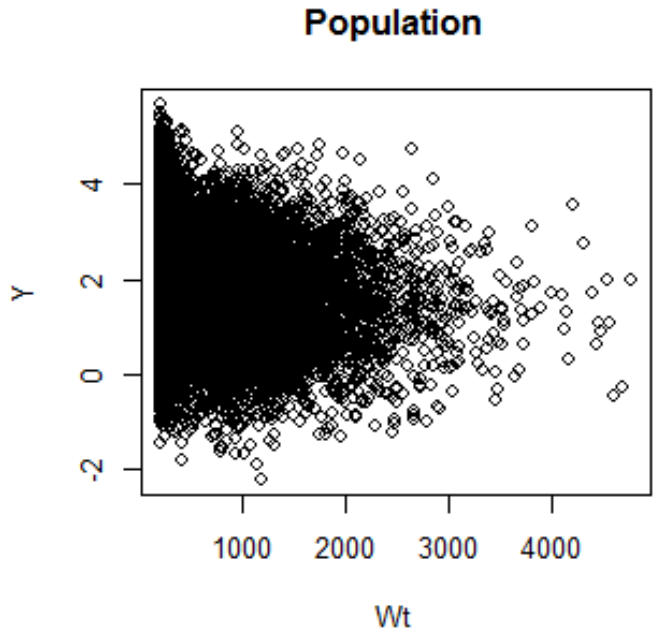


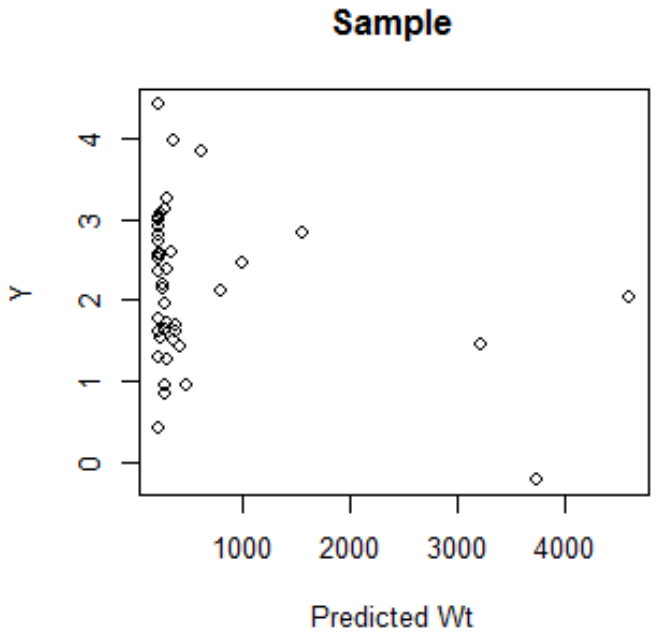
Scenario 6

- Population:
  - N = 20,000
  - $Z_1 \sim$ Bernoulli(0.5)
  - $Z_2|Z_1 \sim$ Bernoulli(0.3 + 0.2Z_1)
  - $U|Z_1, Z_2 \sim N(1 + 0.5Z_1 + 0.3Z_2, 0.01)$
  - $Y|U \sim N(1 + 0.5U + 0.1U^2, 1)$
  - $R|U \sim$ Bernoulli(logit^{-1}(-6.5 + 5U))
- Sample:
  - n = 100, 500
  - SRS
  - Y observed only when R=1
- Weights:
  - Estimate response probability $\hat{\pi}_i$ from logistic regression on U.
  - Weight $w_i = N / n\hat{\pi}_i$

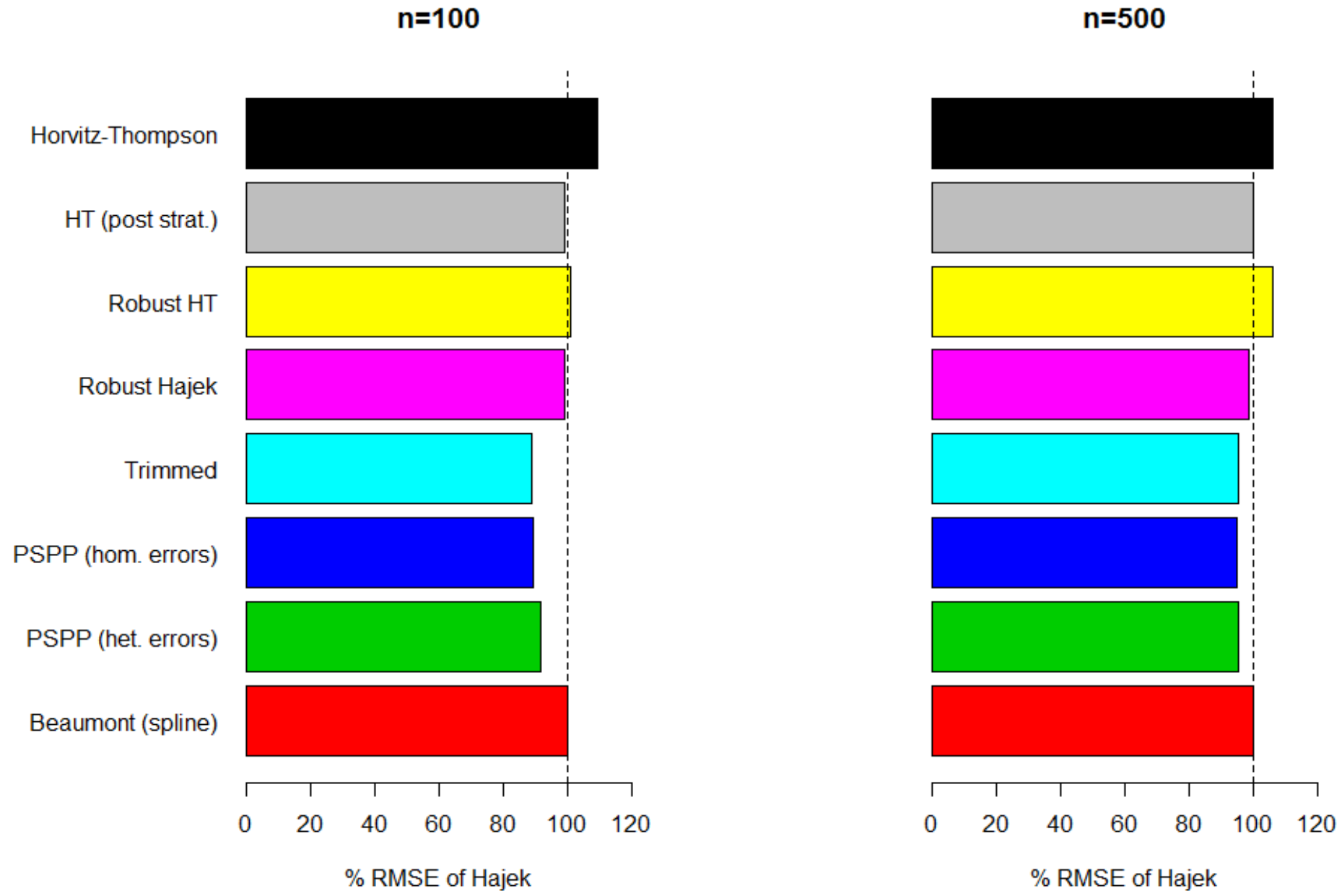

Population
Y
U
-2
0
2
4
1.0
1.5
2.0


Sample
Y
U
0
1
2
3
4
0.8
1.0
1.2
1.4
1.6
1.8


Population
Y
Wt
-2
0
2
4
1000
2000
3000
4000

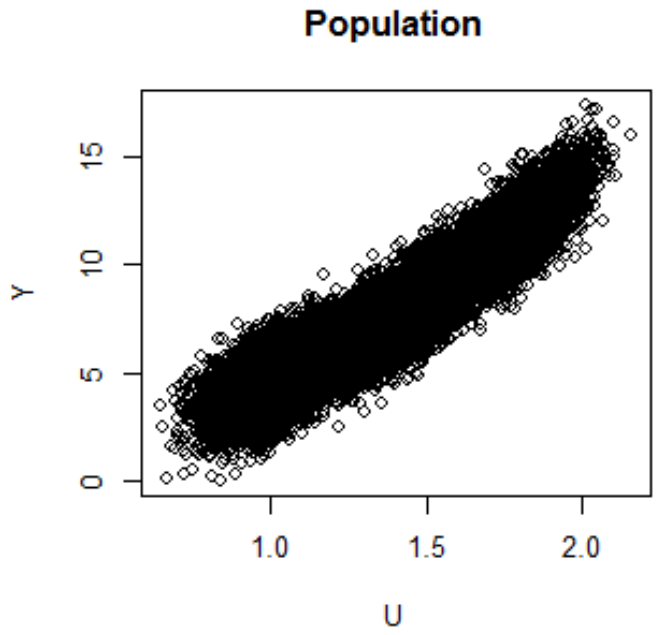

Sample
Y
Predicted Wt
0
1
2
3
4
1000
2000
3000
4000

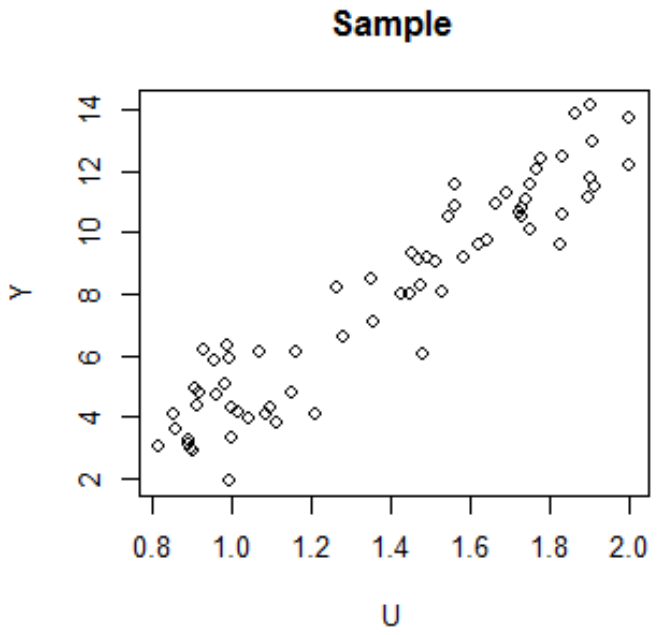
Relative RMSE for Scenario 6

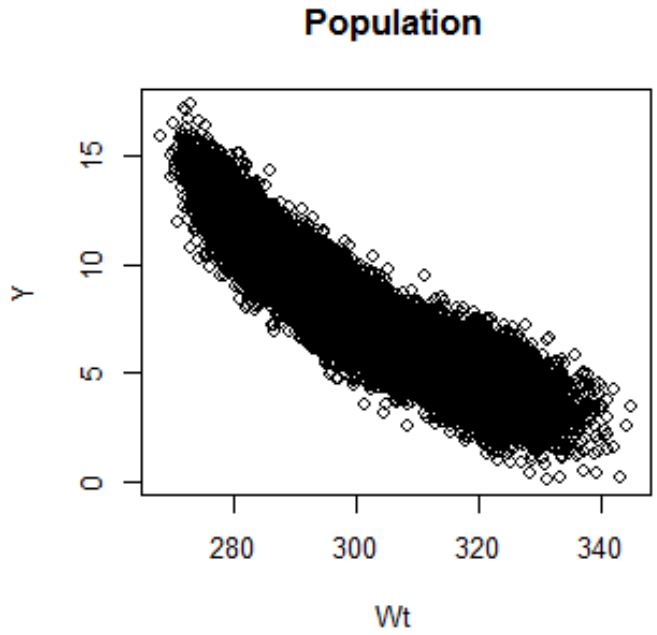


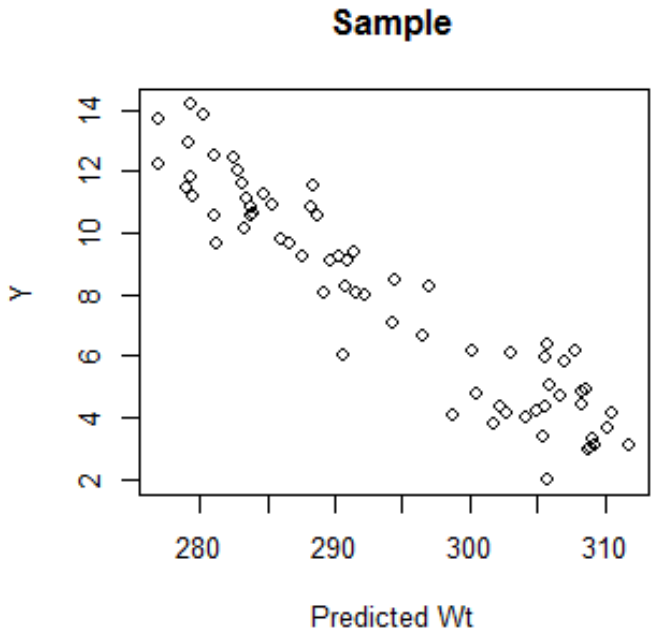
Scenario 7

- Population:
  - N = 20,000
  - $Z_1 \sim$ Bernoulli(0.5)
  - $Z_2|Z_1 \sim$ Bernoulli(0.3 + 0.2Z_1)
  - $U|Z_1, Z_2 \sim N(1 + 0.5Z_1 + 0.3Z_2, 0.01)$
  - $Y|U \sim N(1 + 0.5U + 3U^2, 1)$
  - $R|U \sim$ Bernoulli(logit^{-1}(0.5U))
- Sample:
  - n = 100, 500
  - SRS
  - Y observed only when R=1
- Weights:
  - Estimate response probability $\hat{\pi}_i$ from logistic regression on U.
  - Weight $w_i = N / n\hat{\pi}_i$

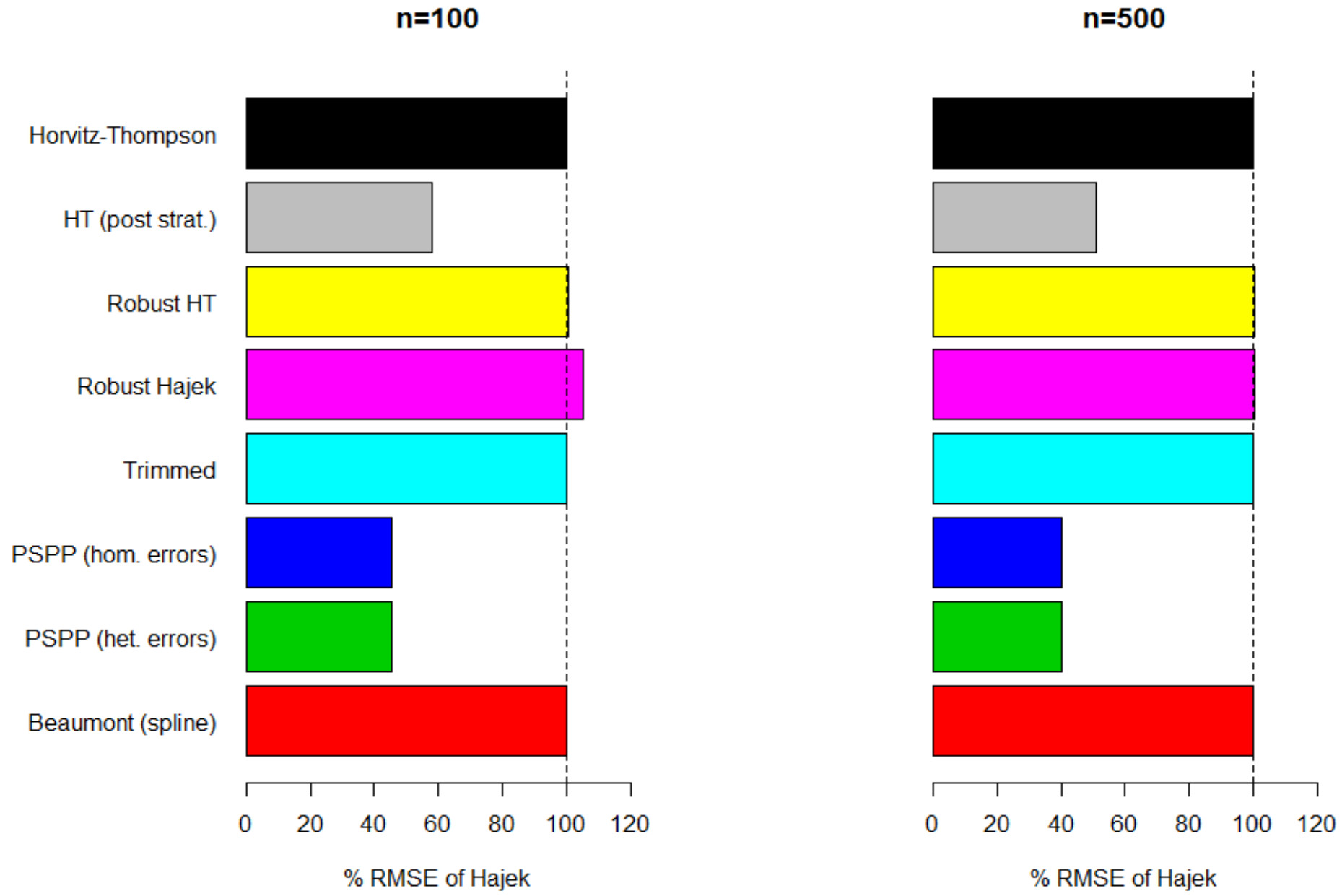

Population
Y
U
0
5
10
15
1.0
1.5
2.0


Sample
Y
U
2
4
6
8
10
12
14
0.8
1.0
1.2
1.4
1.6
1.8
2.0


Population
Y
Wt
0
5
10
15
280
300
320
340

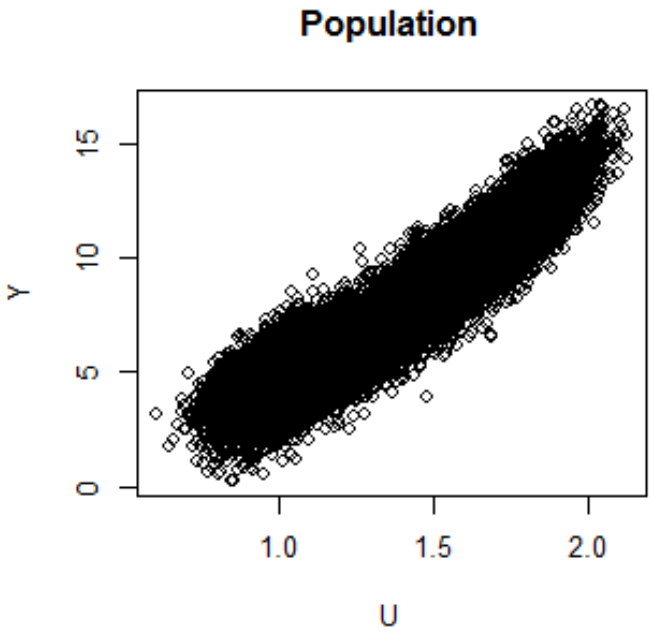

Sample
Y
Predicted Wt
2
4
6
8
10
12
14
280
290
300
310

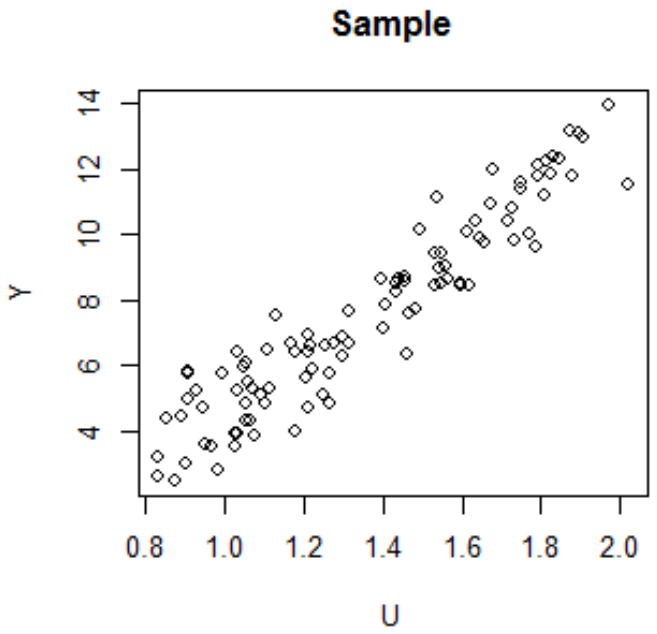
Relative RMSE for Scenario 7

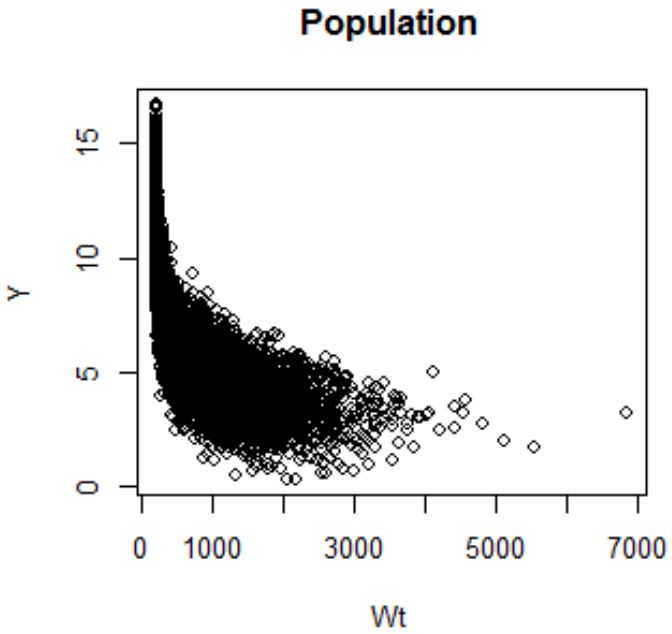


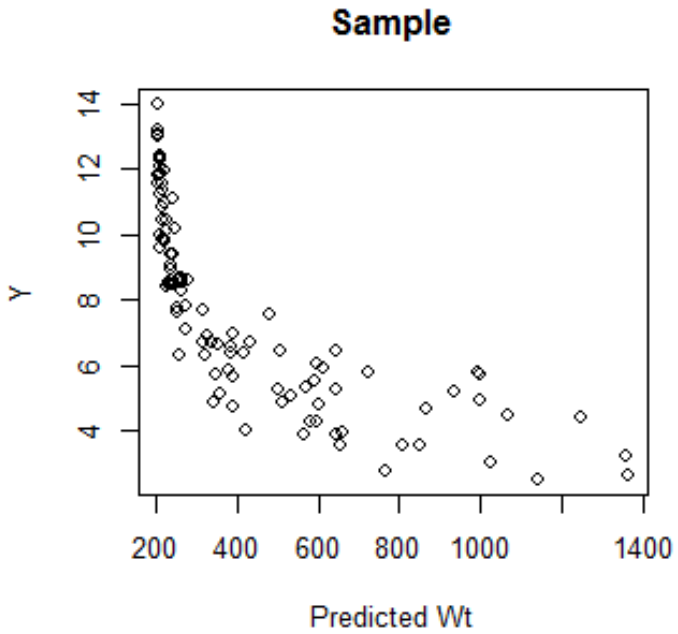
Scenario 8

- Population:
  - N = 20,000
  - $Z_1 \sim$ Bernoulli(0.5)
  - $Z_2|Z_1 \sim$ Bernoulli(0.3 + 0.2Z_1)
  - $U|Z_1, Z_2 \sim N(1 + 0.5Z_1 + 0.3Z_2, 0.01)$
  - $Y|U \sim N(1 + 0.5U + 3U^2, 1)$
  - $R|U \sim$ Bernoulli(logit^{-1}(-6.5 + 5U))
- Sample:
  - n = 100, 500
  - SRS
  - Y observed only when R=1
- Weights:
  - Estimate response probability $\hat{\pi}_i$ from logistic regression on U.
  - Weight w_i = N / $n\hat{\pi}_i$

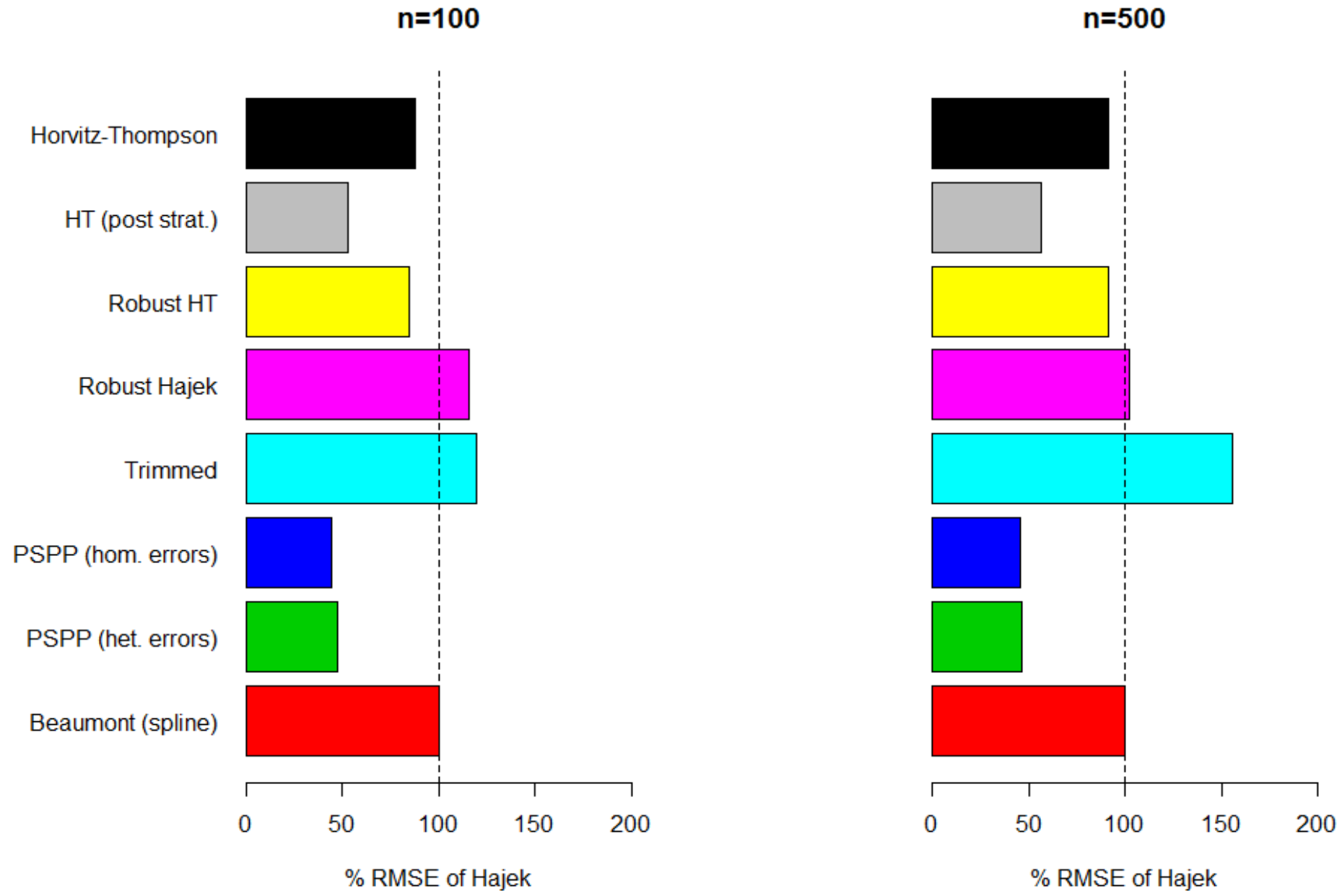

Population
Y
U
0
5
10
15
1.0
1.5
2.0


Sample
Y
U
4
6
8
10
12
14
0.8
1.0
1.2
1.4
1.6
1.8
2.0


Population
Y
Wt
0
5
10
15
0
1000
3000
5000
7000

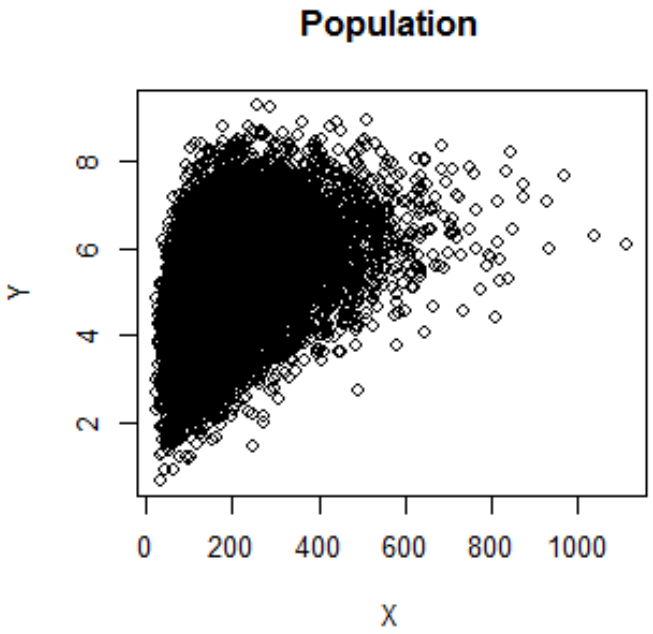

Sample
Y
Predicted Wt
4
6
8
10
12
14
200
400
600
800
1000
1400

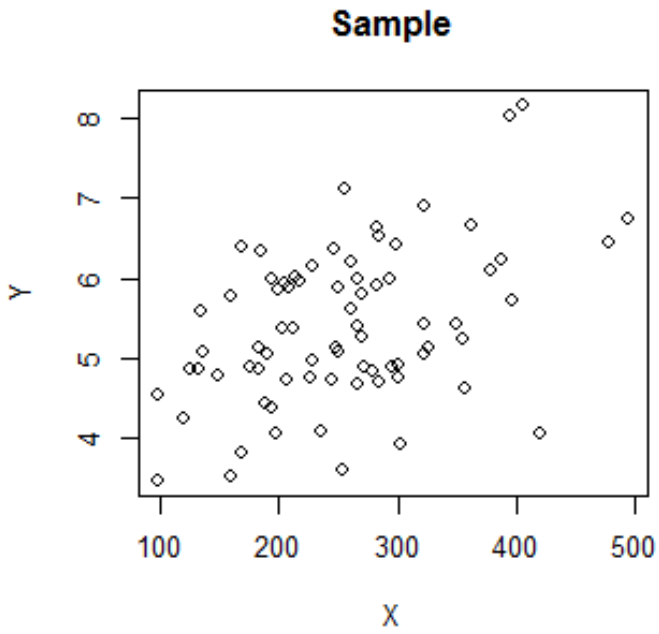
Relative RMSE for Scenario 8

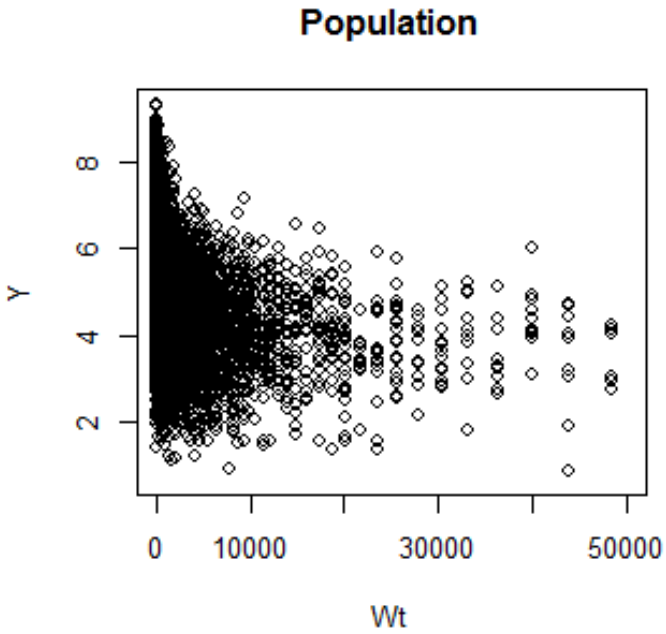


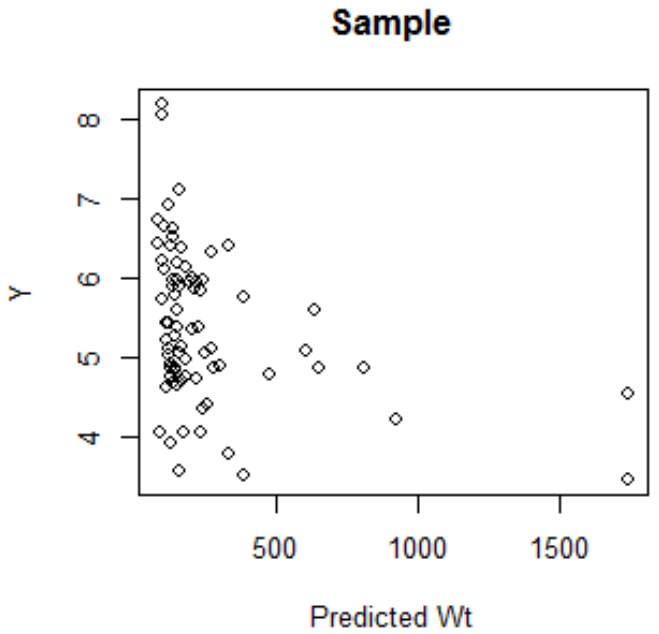
Scenario 9

- Population:
  - N = 20,000
  - $X|Z = \text{floor}(100Z)$, $\log(Z) \sim N(0.5, 0.5)$
  - $Y|X \sim N(10 + \log X + 2\log^2 X, \log^2 X)$
  - $R|X \sim \text{Bernoulli}(\text{logit}^{-1}(-15 + 3\log X))$
- Sample:
  - n = 100, 500
  - PPS sampling with X as size variable
  - Y observed only when R=1
- Weights:
  - Estimate response probability $\hat{\pi}_i$ from logistic regression on X.
  - Weight $w_i = N / n\hat{\pi}_i$

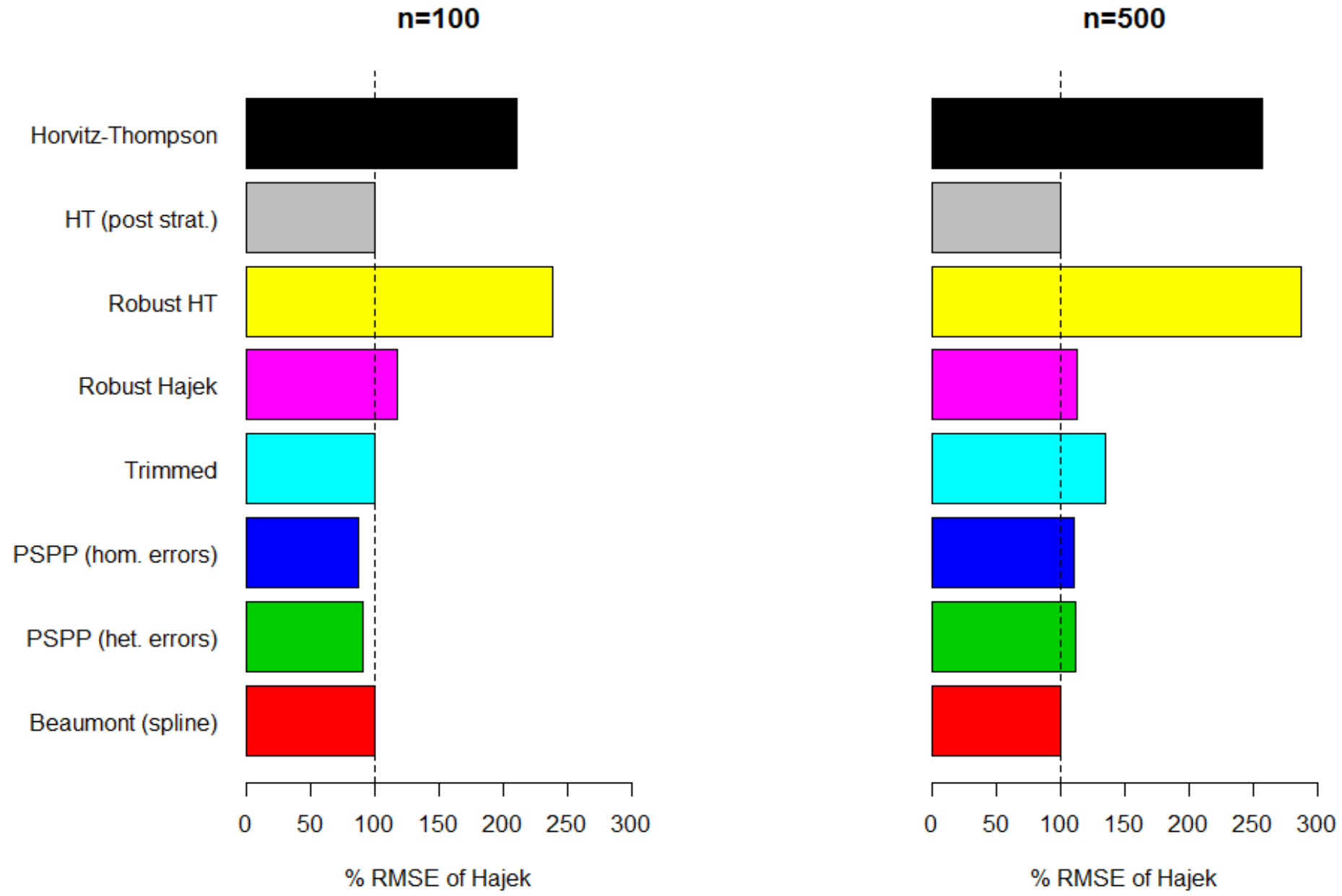

Population
Y
X
0
200
400
600
800
1000
2
4
6
8


Sample
Y
X
100
200
300
400
500
4
5
6
7
8


Population
Y
Wt
0
10000
30000
50000
2
4
6
8

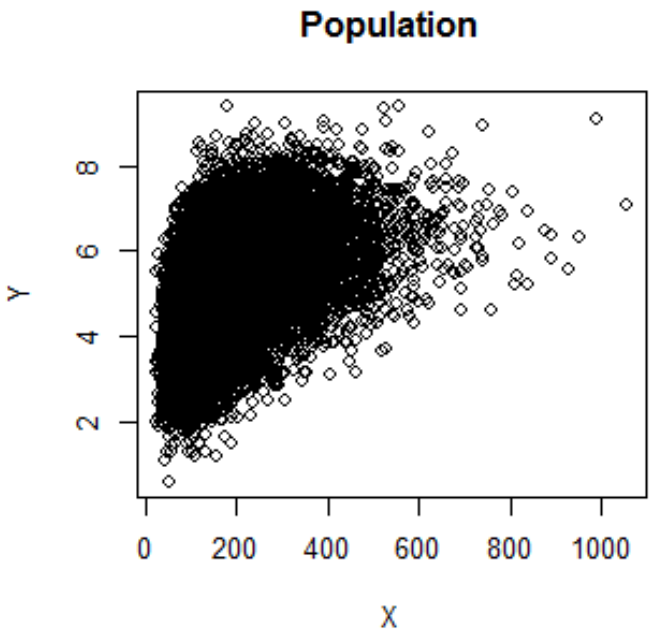

Sample
Y
Predicted Wt
500
1000
1500
4
5
6
7
8

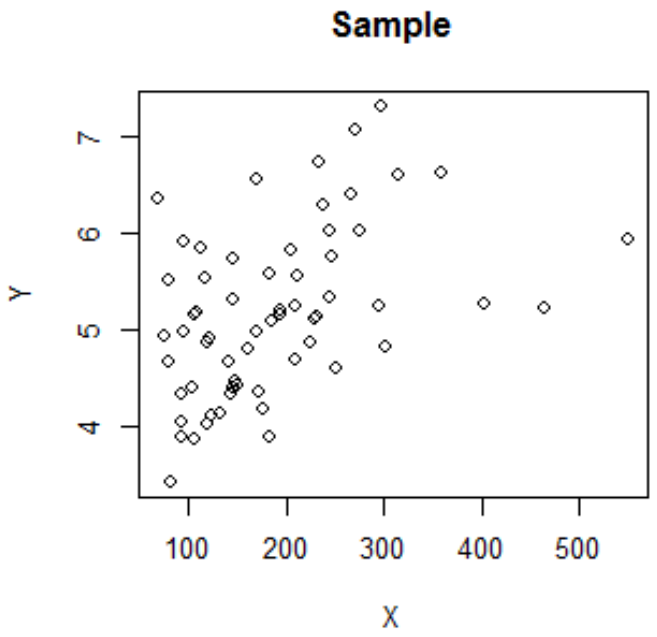
Relative RMSE for Scenario 9

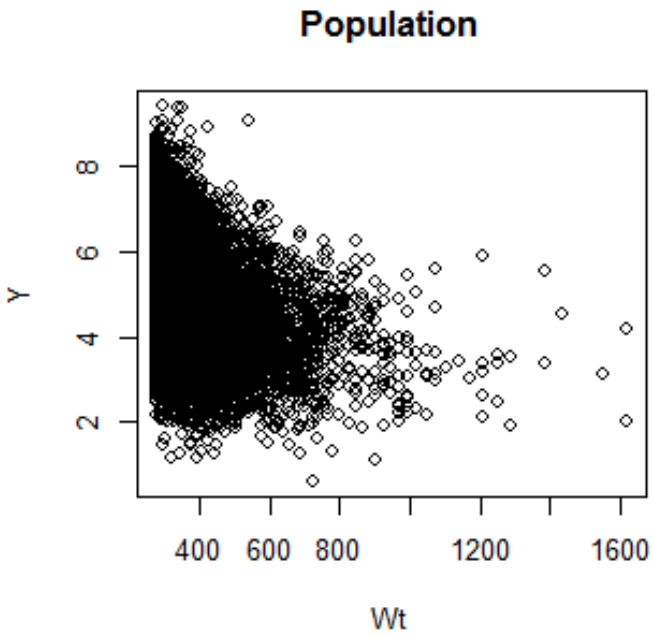


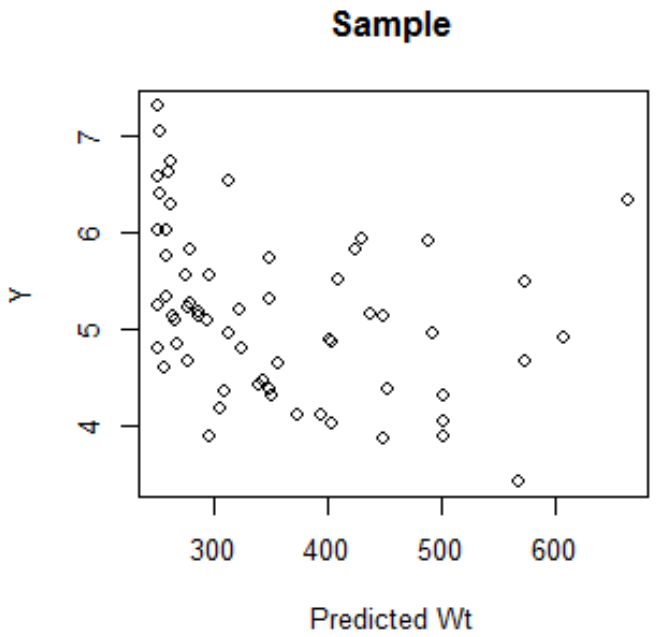
Scenario 10

- Population:
  - N = 20,000
  - $X|Z = \text{floor}(100Z)$, $\log(Z) \sim N(0.5, 0.5)$
  - $Y|X \sim N(10 + \log X + 2\log^2 X, \log^2 X)$
  - $R|X \sim \text{Bernoulli}(\text{logit}^{-1}(11.2 - 2\log X))$
- Sample:
  - n = 100, 500
  - PPS sampling with X as size variable
  - Y observed only when R=1
- Weights:
  - Estimate response probability $\hat{\pi}_i$ from logistic regression on X.
  - Weight $w_i = N / n\hat{\pi}_i$

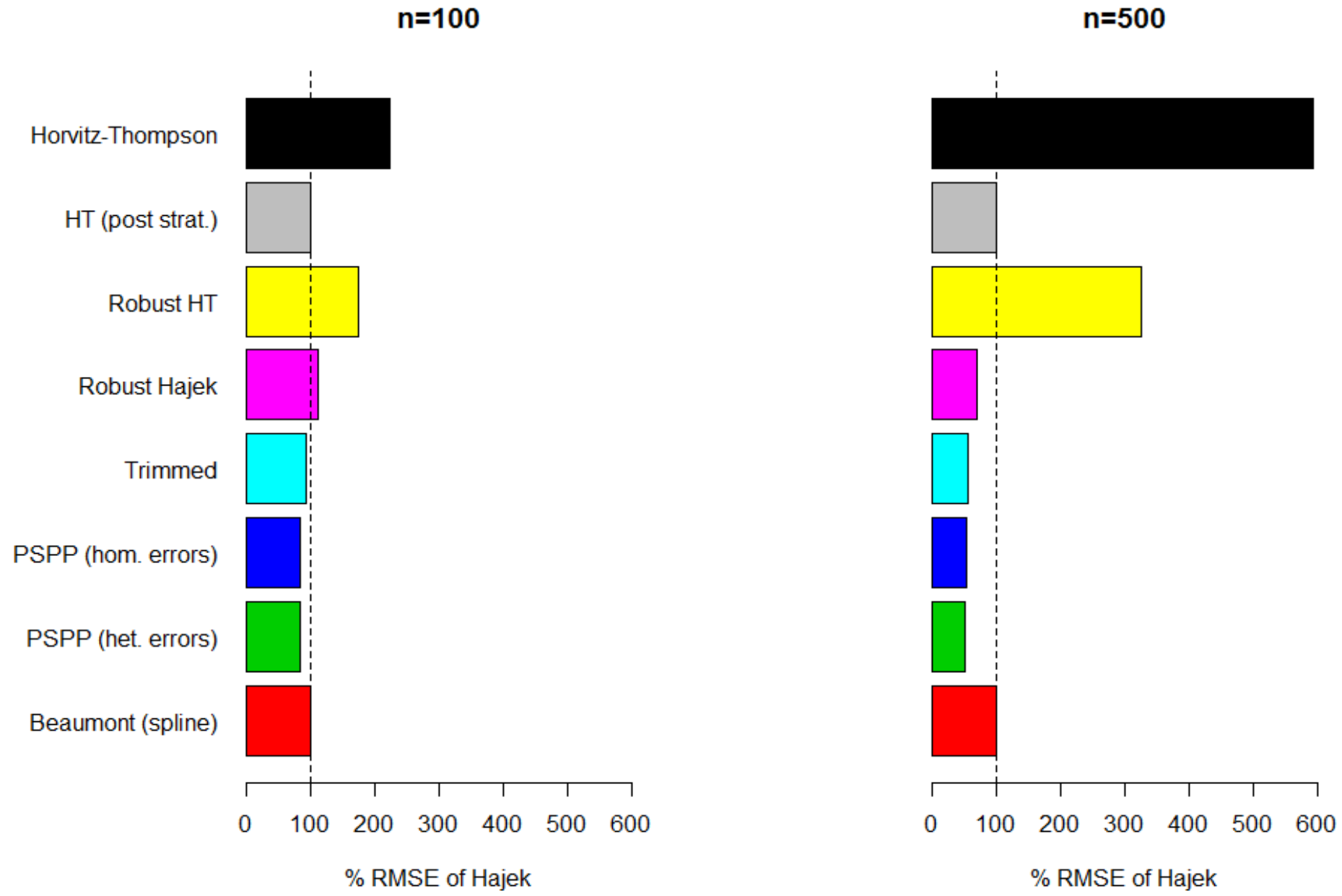

Population
Y
X
0
200
400
600
800
1000
2
4
6
8


Sample
Y
X
100
200
300
400
500
4
5
6
7


Population
Y
Wt
400
600
800
1200
1600
2
4
6
8


Sample
Y
Predicted Wt
300
400
500
600
4
5
6
7

Relative RMSE for Scenario 10



Summary

- Hajek generally better than Horvitz-Thompson when y_i and π_i are not proportional.
- Weight trimming fails when y_i is strongly associated with w_i.
- Little to no difference between Beaumont and Hajek.
- PSPP generally does best when y is a continuous function of π, though less successful in discontinuous functions and in presence of extreme outliers.
- Minor differences between variance structures in PSPP, with small gains when variance of error is not constant.

Next steps

- Explore additional models with discontinuous mean functions and extreme outliers.

- Variance estimation and inference.