Assessment of standard and alternative linear regression estimators

Sung Chul (David) Hong, Jiaxuan (Jessie) Liu, Xin Xu, Hao Xue

HARVARD T.H. CHAN SCHOOL OF PUBLIC HEALTH

Introduction

- Ordinary least squares (OLS) estimator is the standard method for linear regression that minimizes the sum of squared residuals. However, it assumes homoscedasticity. By the Gauss-Markov theorem, OLS is the best linear unbiased estimator (BLUE) under homoscedastic errors.
- Weighted least squares (WLS) estimator minimizes the sum of squared residuals with optimal weights which are inverse of OLS-fitted residuals (no homoscedasticity assumption). Under heteroscedasticity, WLS is BLUE.
- Least absolute deviation (LAD) regression estimates the conditional median of a dependent variable given the independent variable(s) by minimizing sums of absolute deviations between observed and predicted values. It is more robust to outliers and is more efficient for heavy-tailed error distributions.
- In this project, we assessed the performance of OLS, WLS, and LAD estimators under three different error settings.

Table 1. Summary of three estimators

Method	Algorithm	Point Estimator $\hat{\beta}$	Variance estimator $Var(\hat{\beta})$
OLS	$\min \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$	$(X^T X)^{-1} X^T Y$	$\hat{\sigma}^2 (X^T X)^{-1}$ $\hat{\sigma}^2$=MSE
WLS	$\min \sum_{i=1}^{n} \omega_i (y_i - \hat{y}_i)^2$	$(X^T \hat{\Sigma}^{-1} X)^{-1} X^T \hat{\Sigma}^{-1} Y$ $\hat{\Sigma} = diag(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots \hat{\sigma}_n^2)$	$(X^T \hat{\Sigma}^{-1} X)^{-1}$ $\hat{\Sigma} = diag(\hat{\sigma}_1^2, \hat{\sigma}_2^2, \ldots \hat{\sigma}_n^2)$
LAD	$\min \sum_{i=1}^{n} \lvert y_i - \hat{y}_i \rvert$	No closed form, estimated by simplex method	No closed form, estimated by bootstrap

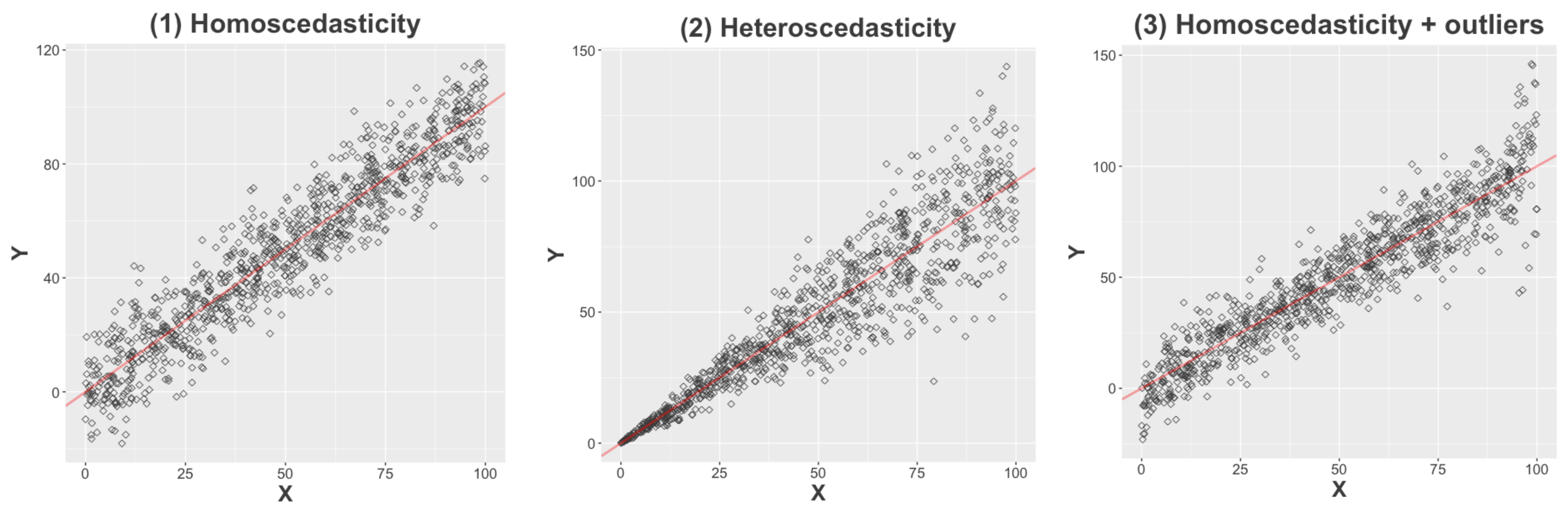


Figure 1. Simulated data under three error settings

Methods

1. Simulation

- We randomly generated 1000 X_i from Uniform [0,100].
- $Y_i = X_i + \varepsilon_i$ (true β_0 =0, β_1 =1)
- Three different settings of ε_i were simulated:

(i) Homoscedasticity: $\varepsilon_i \sim$N(0, 10)

(ii) Heteroscedasticity: $\varepsilon_i \sim$ N(0, X_i)

(iii) Homoscedasticity + outliers:

- for $X_i \leq 95$, $\varepsilon_i \sim$N(0, 10)
- for $X_i > 95$, $\varepsilon_i \sim$N(5, 0.3$*X_i$)

- 100 datasets were simulated under each error setting
- An example dataset of simulated data is shown in Figure 1

2. Point estimation and confidence interval (95% CI) for β_1

- All regressions were conducted in R:
  - OLS: lm(y~x)
  - WLS: lm(y~x, weights)
  - LAD: lad(y~x)
- For OLS and WLS, we used the Wald 95% CI generated from the lm() function. For LAD, we estimated 95% CI by bootstrapping.

3. Estimator assessment

- Distribution of $\hat{\beta}_1$ (unbiasedness and efficiency)
- 95% CI width under different sample size
- 95% CI coverage probability: we repeated the simulation for 100 iterations, and average coverage probability was calculated for each iteration.

Results

1. Distribution of $\hat{\beta}_1$ under different error settings

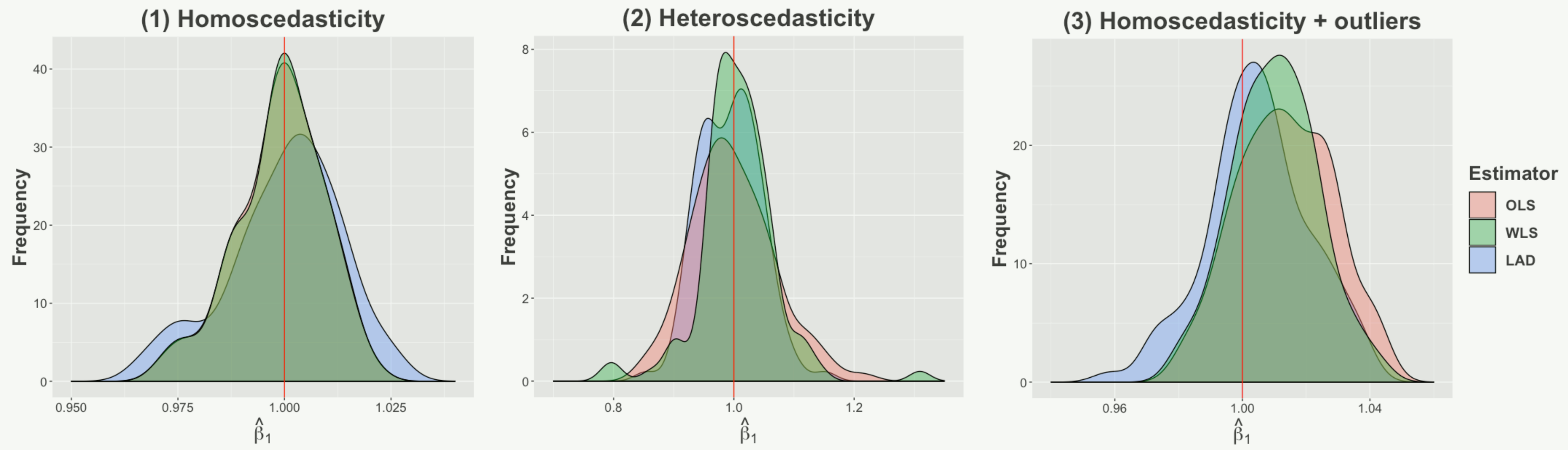


Figure 2.
(1) Homoscedasticity: OLS and WLS have similar distributions, both of which are centered at the true β_1 =1 (red line); the center of LAD is positively biased.
(2) Heteroscedasticity: WLS is best centered at 1, and has the smallest spread.
(3) Homoscedasticity + outliers: LAD is the least biased while OLS is the most biased.

2. 95%CI width across different same sizes

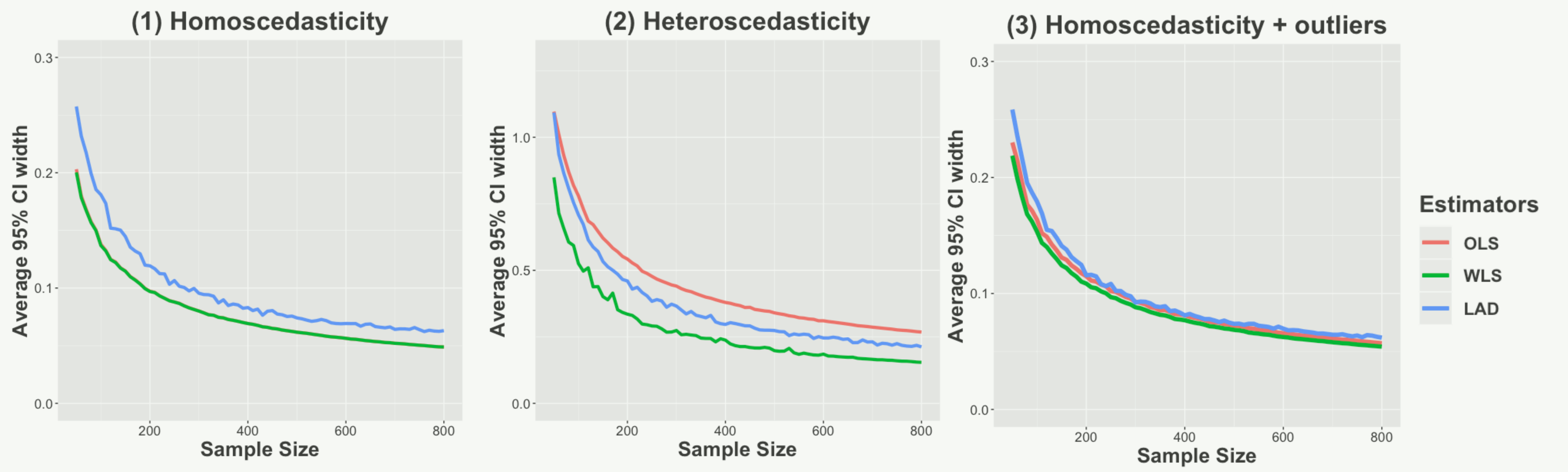


Figure 3.
95% CI width decrease with sample size for all settings.
Consistent with our findings in Figure 2:
(1) Homoscedasticity: OLS and WLS have similar width; LAD has a generally wider CI.
(2) Heteroscedasticity: WLS has the smallest width, while OLS has the largest.
(3) Homoscedasticity + outliers: all three estimators have similar width.

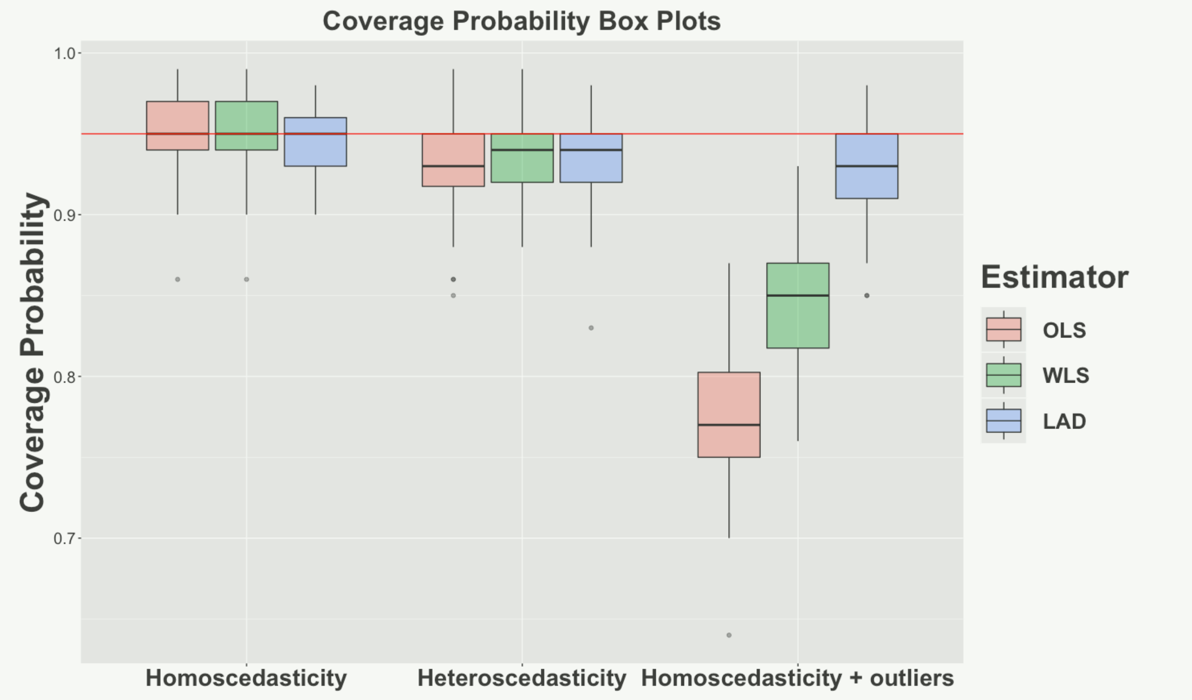
3. Coverage probability of 95% CI

Figure 4. Distribution of 95% CI coverage probability across 100 iterations

(1) Homoscedasticity: all three estimators have mean coverage probability around 0.95, while LAD is slightly lower than the other two; OLS and WLS have similar coverage, considering their similar distribution under homoscedastic errors (Figure 2&3);

(2) Heteroscedasticity: the average mean probability of three estimators are all slightly below 0.95; WLS yielded better coverage than OLS by accounting for heteroscedasticity; LAD also performed better as it is less sensitive to skewed errors;

(3) Homoscedasticity + outliers: the average coverage probability of LAD is close to 0.95, while that for the other two are much lower than 0.95, with OLS being the lowest.



Conclusion

- Our results are consistent with Gauss-Markov theorem, that OLS and WLS are BLUE under homoscedastic and heteroscedastic errors, respectively.
- LAD is recommended when having outliers in data. However, LAD does not perform well with homoscedastic errors.

Future directions

We only explored a few specific settings for this assessment due to limited time for this project. However, we came up with many interesting directions for future investigation, including:

- Experiment the efficacy of WLS under different heteroscedastic error settings;
- Examine the power of the estimators by computing the probability of rejecting the null when the alternative is true across different values of β_1;
- Increase the number of iterations in estimating 95%CI coverage probabilities.