MATH-517: Assignment 3

Robertson Rahul

2025-04-10

Theoretical exercise

We have that $\hat{\beta} = (X^TWX)^{-1}X^TWY$ so it follows that we can express β_0 as a linear combination of the $Y_i's$ that only depends on the elements that are in X and W, that is x, X_i, K, h §

Step 1: Compute ((X^T W X))

$$X^TWX = \begin{bmatrix} 1 & X_1 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{bmatrix}^T \begin{bmatrix} K\left(\frac{X_1-x}{h}\right) & & \\ & \ddots & \\ & & K\left(\frac{X_n-x}{h}\right) \end{bmatrix} \begin{bmatrix} 1 & X_1 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{bmatrix}$$

$$= \begin{bmatrix} 1 & X_1 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{bmatrix}^T \begin{bmatrix} K\left(\frac{X_1-x}{h}\right) & (X_1 - x)K\left(\frac{X_1-x}{h}\right) \\ \vdots & \vdots \\ K\left(\frac{X_n-x}{h}\right) & (X_n - x)K\left(\frac{X_n-x}{h}\right) \end{bmatrix}$$

$$= \begin{bmatrix} \sum_{i=1}^n K\left(\frac{X_i-x}{h}\right) & \sum_{i=1}^n (X_i - x)K\left(\frac{X_i-x}{h}\right) \\ \sum_{i=1}^n (X_i - x)K\left(\frac{X_i-x}{h}\right) & \sum_{i=1}^n (X_i - x)^2 K\left(\frac{X_i-x}{h}\right) \end{bmatrix}$$

$$= nh \begin{bmatrix} S_{n,0}(x) & S_{n,1}(x) \\ S_{n,1}(x) & S_{n,2}(x) \end{bmatrix}$$

Since we have a 2 by 2 matrix, we know how to calculate it's inverse

$$(X^TWX)^{-1} = \frac{1}{(nh)^2 S_{n,2}(x)S_{n,0}(x) - S_{n,1}(x)^2} \begin{bmatrix} S_{n,2}(x) & -S_{n,1}(x) \\ -S_{n,1}(x) & S_{n,0}(x) \end{bmatrix}$$

To simplify notations we write $K(\frac{X_i-x}{h}) = K_i$ and $\frac{1}{(nh)^2 S_{n,2}(x)S_{n,0}(x)-S_{n,1}(x)^2} = D\_n$

$$(X^TWX)^{-1}X^TW = \frac{1}{D_n}\begin{bmatrix} S_{n,2}(x) & -S_{n,1}(x) \\ -S_{n,1}(x) & S_{n,0}(x) \end{bmatrix}\begin{bmatrix} K_1 & \dots & K_n \\ (X_1-x)K_1 & \dots & (X_n-x)K_n \end{bmatrix}$$

$$= \frac{1}{D_n}\begin{bmatrix} S_{n,2}(x)K_1 - S_{n,1}(x)(X_1-x)K_1 & \dots & S_{n,2}(x)K_n - S_{n,1}(x)(X_n-x)K_n \\ -S_{n,1}(x)K_1 + S_{n,0}(x)(X_1-x)K_1 & \dots & -S_{n,1}(x)K_n + S_{n,0}(x)(X_n-x)K_n \end{bmatrix}$$

The w_{ni} correspond to the i'th coordinate on the first row of this matrix so we find that:

$$w_{ni}(x) = \frac{1}{nh} \cdot \frac{K\left(\frac{x-X_i}{h}\right)\left(S_{n,2}(x) - (X_i-x)S_{n,1}(x)\right)}{S_{n,0}(x)S_{n,2}(x) - S_{n,1}(x)^2}$$

3) Let

$$D_n = \frac{1}{nh\, S_{n,0}(x)S_{n,2}(x) - S_{n,1}^2(x)}.$$

We have

$$\sum_{i=1}^{n} w_{ni}(x) = \frac{\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)S_{n,2}(x) - \sum_{i=1}^{n}(X_i-x)S_{n,1}(x)K\left(\frac{X_i-x}{h}\right)}{D_n}.$$

Since

$$\sum_{i=1}^{n} K\left(\frac{X_i-x}{h}\right) = nh\, S_{n,0}(x) \quad \text{and} \quad \sum_{i=1}^{n}(X_i-x)K\left(\frac{X_i-x}{h}\right) = nh\, S_{n,1}(x),$$

it follows that

$$\sum_{i=1}^{n} w_{ni}(x) = \frac{nh\, S_{n,0}(x)S_{n,2}(x) - nh\, S_{n,1}^2(x)}{D} = \frac{D_n}{D_n} = 1.$$

Practical exercise

Unless stated otherwise, your answer to the practical part should include the following elements:

- Description of the aim of the simulation study.
- Description of the different quantities that intervene in your simulation study. Explain how these quantities (fixed or random) are defined and the reasoning behind the choices you made.
- Description of your findings using appropriate graphics and/or tables (with a caption!) that are well commented in the text.

The code should not appear in the PDF report, unless there is a specific reason to include it.

1. Aim of the Simulation Study

The goal of this simulation study is to explore how the optimal bandwidth h_{amise} and it's estimate using block-wise polynomial estimation $\hat{h}_{amise}$ evolve with respect to the variation of different parameters and hyper-parameters. We also would like to see what number of blocks tends to be the best in order to do such estimations, and weather the number depends on the total amount of data available or not.

2. Quantities in the Simulation

The simulation involves the following quantities

n the number of observations that we simulate, we fix this quantity for each experiment, but we allow it to vary between experiments to see how the quality of our fits, as well as how the optimal number of blocks evolves.

N The number of blocs in which we split our data, again this quantity is fixed at each experiment and can be chosen manually or via selection by minimizing Mallow's C_p but we allow it to vary over different experiments.

X, a vector randomly generated following the Beta(α ,β) distribution with values α β set manually in order to have distributions with different concentrations, we choose X randomly as to simulate real data.

$m(X)$ our regression function $= \sin(\frac{1}{\frac{x}{3}+0.1})$ which we do not change over different experiments.

Y our response variable given by $Y = m(X) + \epsilon$ for $\epsilon \sim \mathcal{N}(0, \sigma^2)$ where we fix σ^2 in order to have a robust analysis.

h_{amise} the true value of the optimal bandwidth given by calculating $\theta_{22} = \int m''(x)^2 f_X(x) dx$ and then setting $h_{amise} = n^{-1/5} \left(\frac{35\sigma^2}{\theta_{22}} \right)^{1/5}$

$\widehat{h}_{amise}$ our estimated value of the optimal bandwidth given by fitting 4th degree polynomials $m_i(x)$ in each block and then estimating θ_{22} by $\widehat{\theta}_{22}(N) = \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{N} \widehat{m}''_j(X_i) \widehat{m}''_j(X_i) \mathbf{1}_{\{X_i \in X_j\}}$

$\widehat{\sigma}^2(N) = \frac{1}{n-5N} \sum_{i=1}^{n} \sum_{j=1}^{N} \left\{ Y_i - \widehat{m}^j(X_i) \right\}^2 \mathbf{1}_{\{X_i \in X_j\}}$

and so we get that $\widehat{h}_{amise} = n^{-1/5} \left(\frac{35\widehat{\sigma}^2}{\widehat{\theta}_{22}} \right)^{1/5}$

3. Simulation Results and Visualizations

We now look at our plots to assess the effect of our parameters on estimation, when not specified we set $\alpha = 1$ and $\beta = 2$

We first look at the effect of sample size over our estimation of $\widehat{h}_{amise}$, we expect an increase in the amount of observations n to improve our estimation and this is indeed what we see as shown in the following graph.
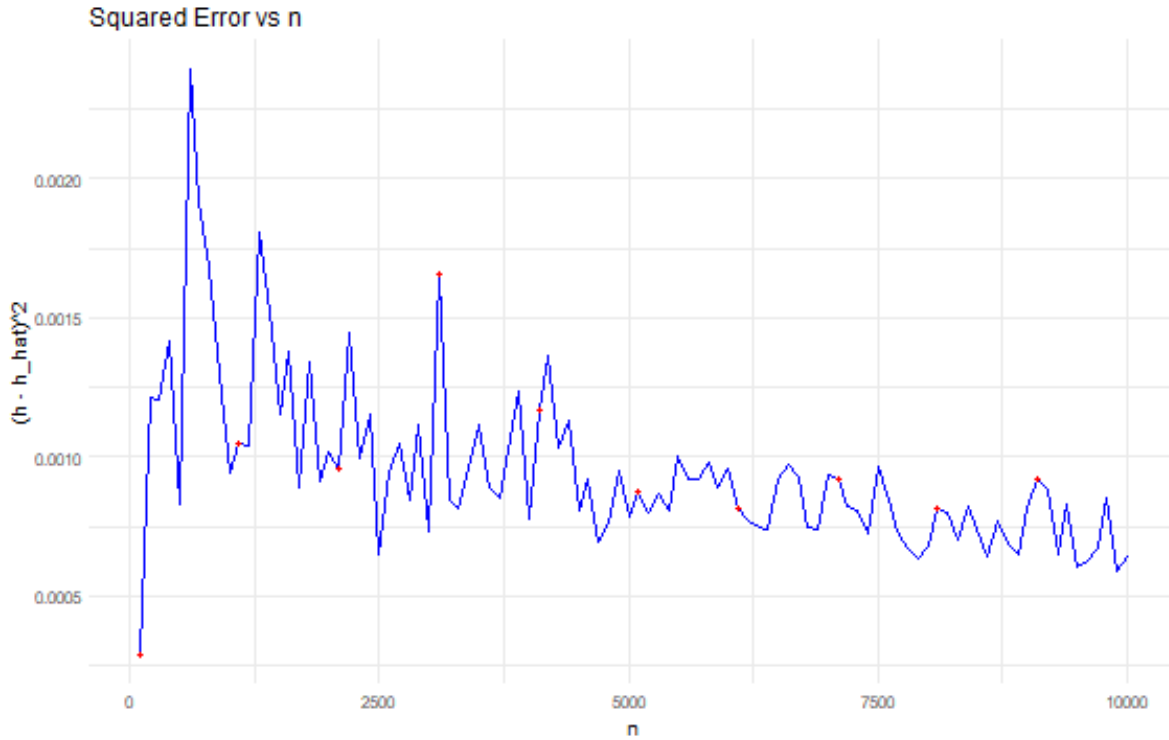

Figure 1: figure 1 : squared error against sample size n

4

Indeed we see that on average the error is decreasing with sample size.

Next the effect of the number of blocks N influences over the estimation of our parameters $\hat{\theta}_{22}$ and $\hat{\sigma}^2$ is shown in the graph below
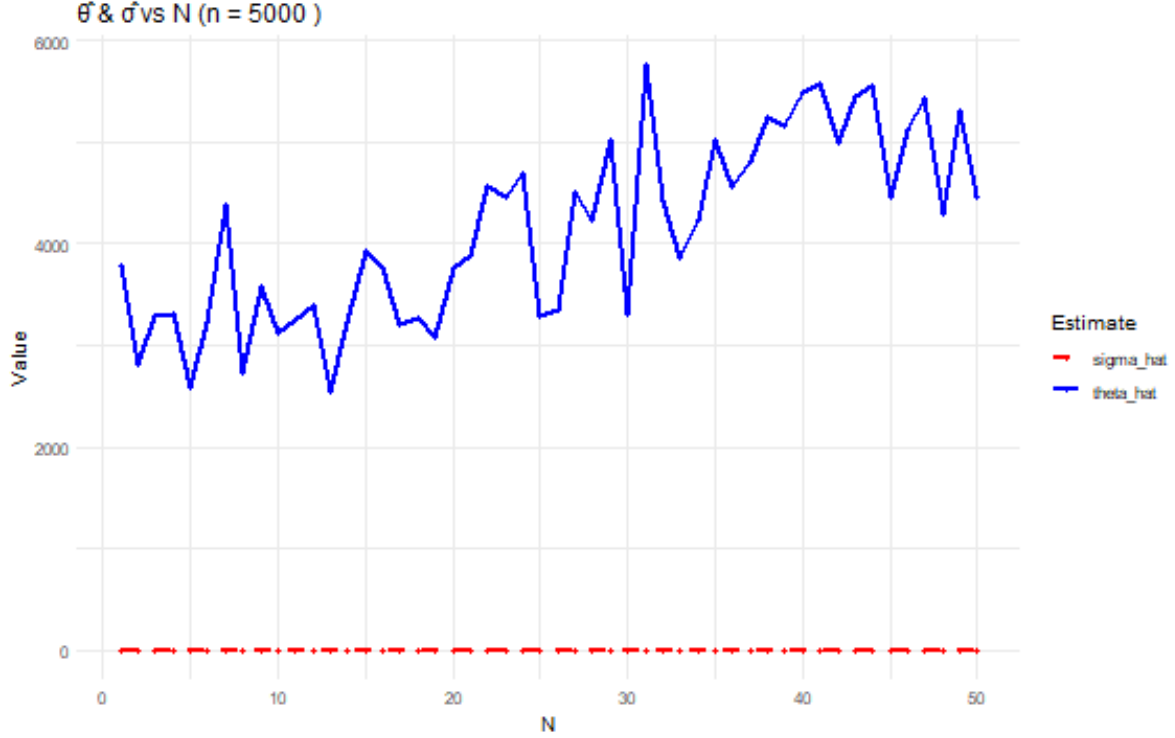

Figure 2: $\hat{\theta}_{22}$ and $\hat{\sigma}^2$ against block size N

We see a steady increase of $\hat{\theta}_{22}$ whilst $\hat{\sigma}^2$ stays stable around the true value. This $\hat{\theta}_{22}$ quantifies the curvature of our polynomial interpretations $m_i(x)$ and so, as N grows, the amount of observations in each blocks gets smaller, and as such the curvature of our fits will increase.

Previously, all our plots have been with $\alpha = 1$ and $\beta = 2$ but now we look at the effect of variation of these parameters on our estimation. Indeed when α is much larger than β and vice versa, the distribution becomes very concentrated in one side of the support, meaning some blocks will get a huge amount of observations whilst others will get very few, this is shown in the histograms below.
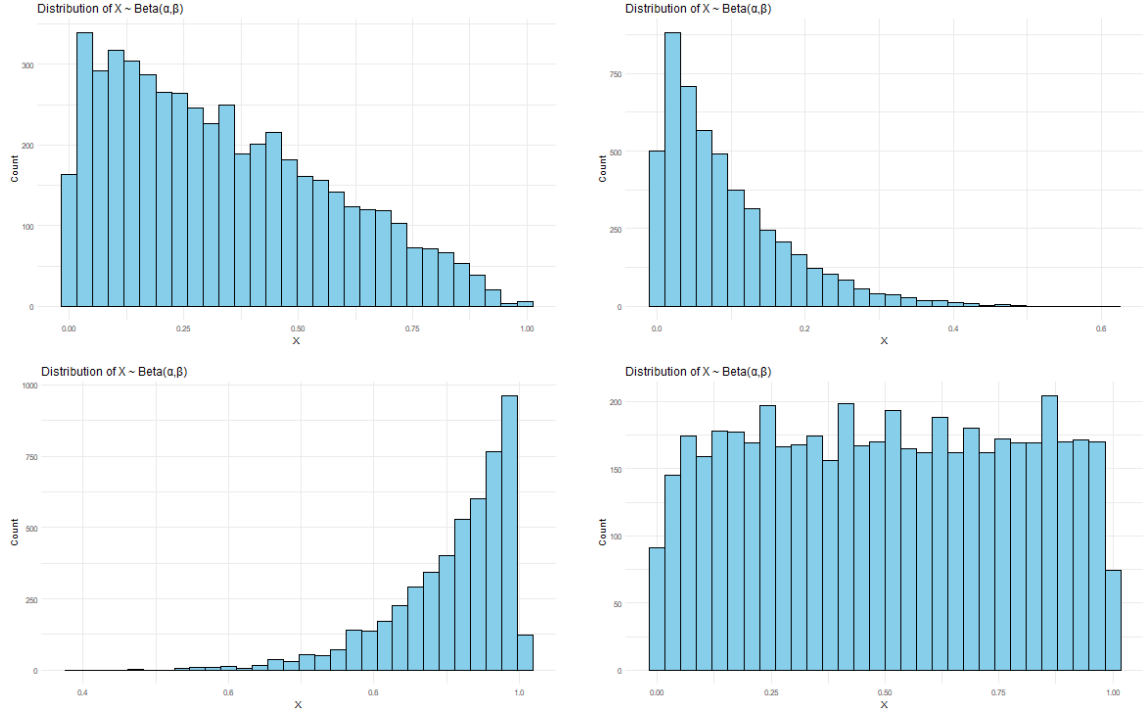
Distributions of Beta(α , β) for, from left to right top to bottom, $\alpha = 1, \beta = 2$, $\alpha = 1, \beta = 10$, $\alpha = 10, \beta = 1$, $\alpha = 1, \beta = 1$

We now assess the quality of fits for each of these distributions ::: {=latex}

:::
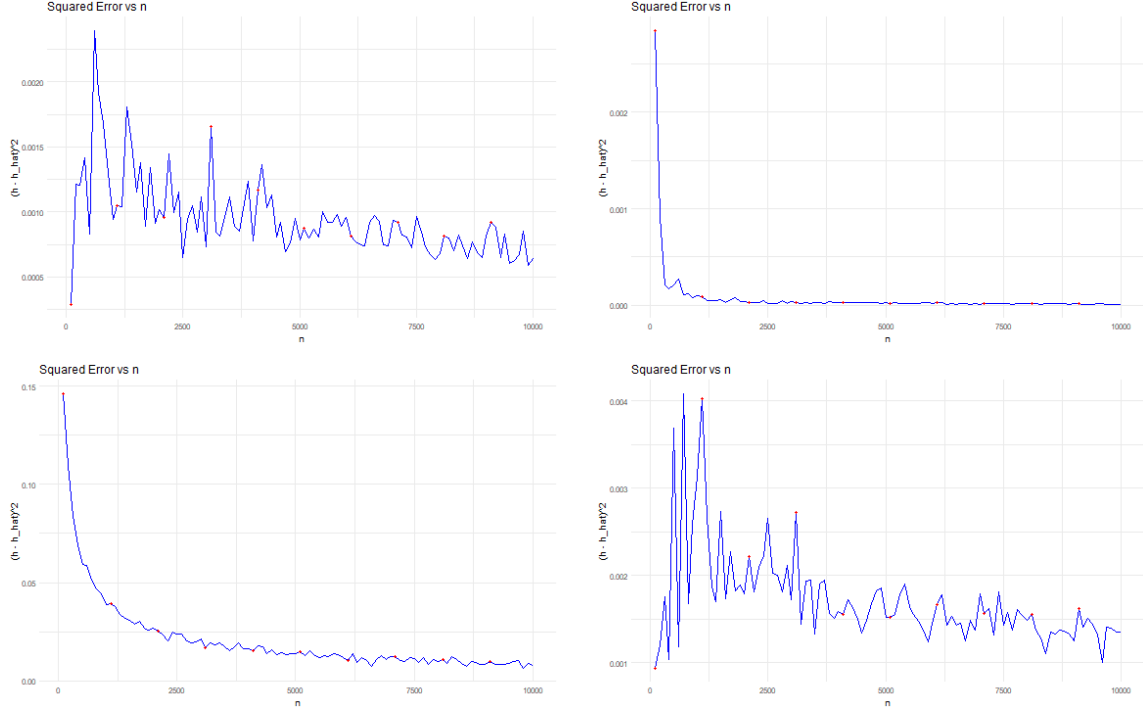
Squared error of h_{amise} for varying parameters of Beta(α , β) for N=10(optimal)

These plots are of great interest. We see that the best plots are in order, alpha =1, beta =10, alpha =1, beta =1, alpha =1, beta =2, alpha =10, beta =1. This is due to the curvature of our regression function m(x). In the case of Beta(1,10), most observations are concentrated near 0, so each block contains many points in a narrow region, allowing the polynomial fits to accurately capture the local behavior of the function $m(x) = sin(1/(x/3 + 0.1))$. For Beta(1,1) (uniform), points are spread evenly across the interval. Although this introduces more variability across blocks, the function is sufficiently smooth in some regions, yielding relatively good fits. For Beta(1,2), the distribution is slightly skewed but less extreme than Beta(1,10), so some blocks cover sparse regions with fewer points, reducing fit accuracy. Finally, for Beta(10,1), most points are concentrated near 1, a region where ( ) m(x) varies rapidly. This causes blocks to have insufficient coverage in regions with high curvature, leading to the poorest fits.

We now interest ourselves in the evolution of $\hat{h}_{amise}$ as N grows, to do this we plot the values of our estimate for different values of N after setting our number of estimations to be 5000
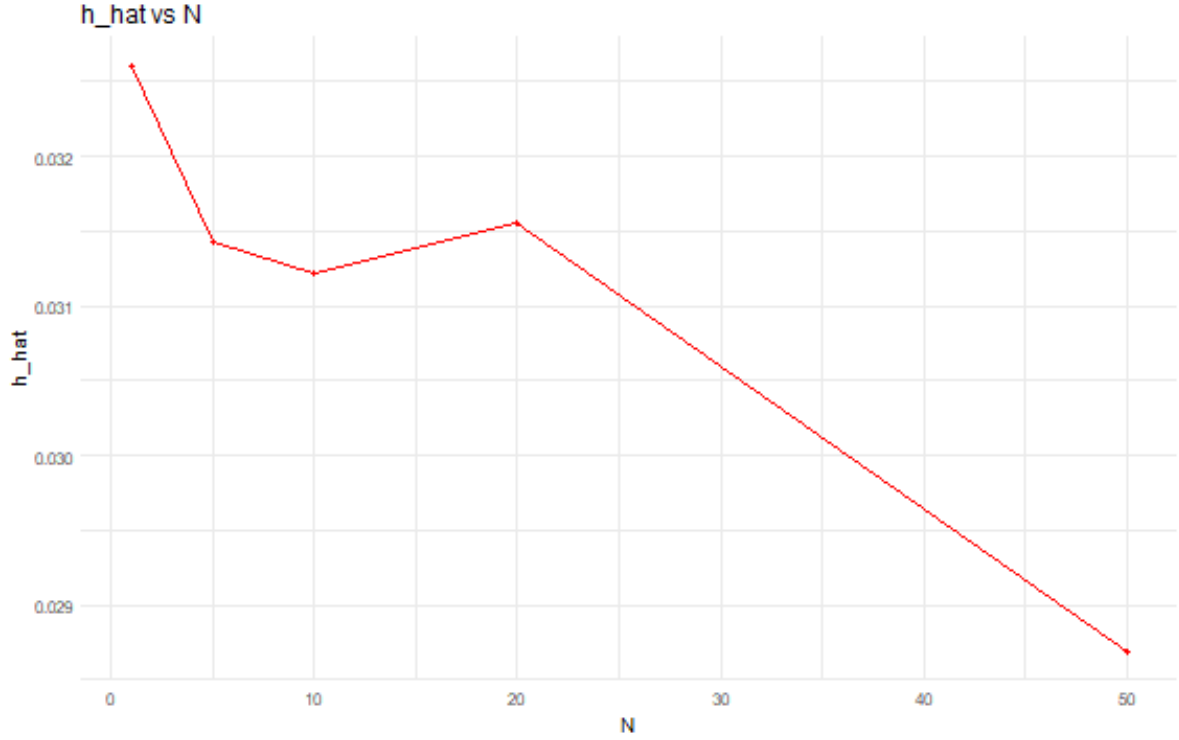
7

Figure 3: evolutioin of estimate value against block number

We see that h_{amise} is decreasing on average and this is exactly what we expect. Indeed we have that the denominator in the computation of h_{amise} is given by

$$n^{1/5}\theta_{22}$$

where θ_{22} is proportional to the curvature of our functions m, and as seen earlier increases with N meaning that the denominator is steadily getting bigger, meaning our estimate will get closer and closer to 0.

We now look into if our choice of block size, should depend on our sample size n. The computation of optimal N is quite expensive, so we were unable to plot the graph of optimal values of N against the sample size n so we instead only look at them for certain points.
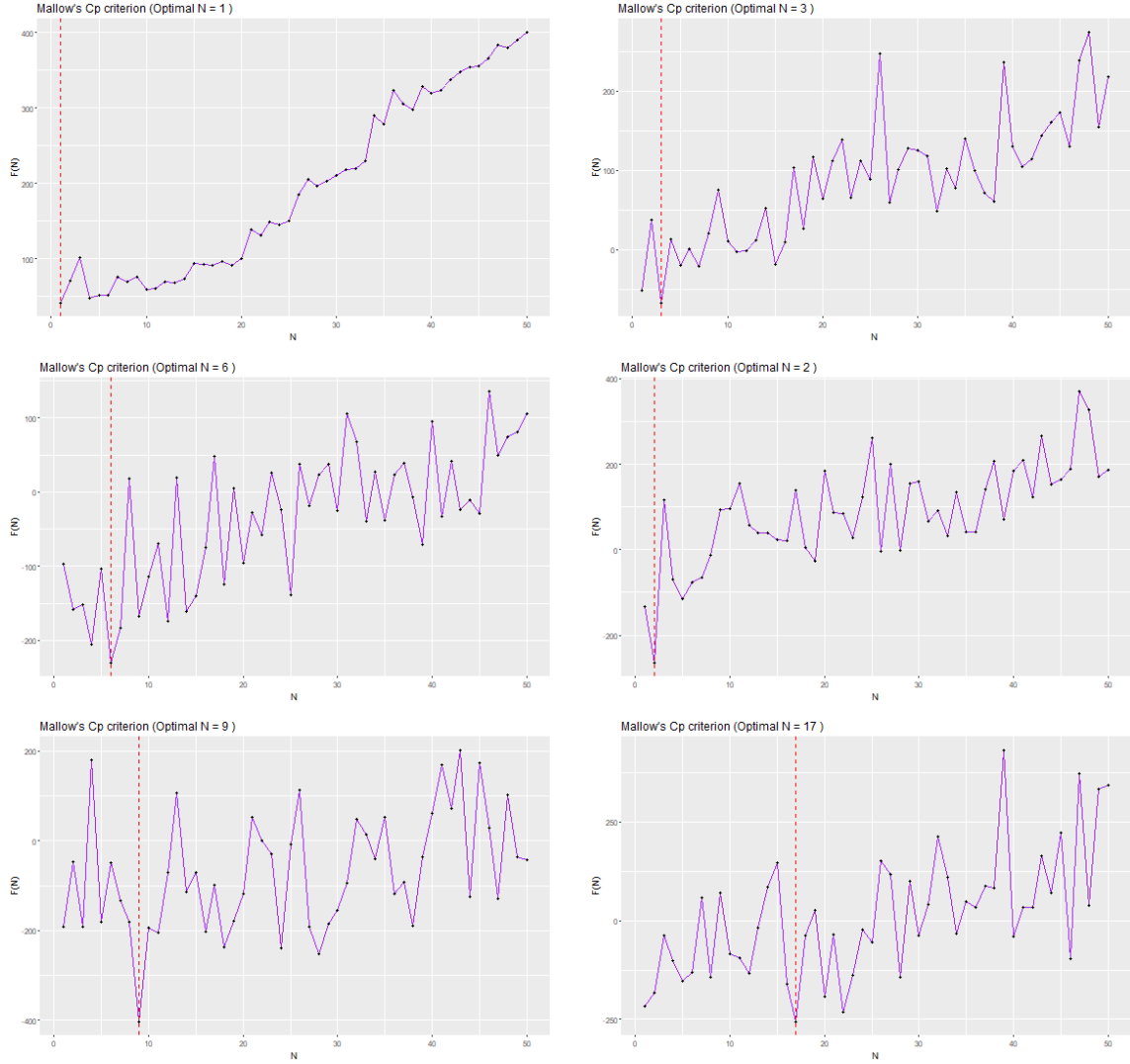
in order left to right top to bottom : n=100 N=1, n=1000 N=3, n=2500 N=6, n=5000 N=2, n=7500 N=9, n=10000 N =17

We see that on average optimal block size grows with number of observations, this makes sense as we expect the number of observations needed per block to stay about the same and so we need more blocks for more observations, this depends on our data however though so can sometimes fail, as we see in the case for n=6000.