MATH-517: Assignment 2

Valerio Viscovo

04/10/2025

Theoretical exercise

1) Proof of the Weighted Least Squares Solution

Let's prove that the solution to the minimization problem

$$\begin{aligned}\widehat{\beta}\,(x) = (\widehat{\beta}_0(x), \widehat{\beta}_1(x)) &= \underset{\beta\in\mathbb{R}^2}{\operatorname{argmin}} \sum_{i=1}^{n} \{Y_i - \beta_0 - \beta_1(X_i - x)\}^2 K\left(\frac{X_i - x}{h}\right) \\ &= \underset{\beta\in\mathbb{R}^2}{\operatorname{argmin}} \left((\mathbf{Y} - \mathbf{X}\beta)^T \mathbf{W}(\mathbf{Y} - \mathbf{X}\beta)\right)\end{aligned}$$

is given by the weighted least squares estimator:

$$\widehat{\beta} = (\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W}\mathbf{Y},$$

where the matrices and vectors are defined as:

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix} \in \mathbb{R}^{n\times 1}, \quad \mathbf{X} = \begin{pmatrix} 1 & X_1 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{pmatrix} \in \mathbb{R}^{n\times 2},$$

$$\mathbf{W} = \operatorname{diag}\left(K\left(\frac{X_1 - x}{h}\right), \ldots, K\left(\frac{X_n - x}{h}\right)\right) \in \mathbb{R}^{n\times n}.$$

The objective function in matrix form is:

$$L(\beta) = (\mathbf{Y} - \mathbf{X}\beta)^t\mathbf{W}(\mathbf{Y} - \mathbf{X}\beta)$$

Expanding the quadratic form yields:

$$L(\beta) = \mathbf{Y}^t\mathbf{W}\mathbf{Y} - \beta^t\mathbf{X}^t\mathbf{W}\mathbf{Y} - \mathbf{Y}^t\mathbf{W}\mathbf{X}\beta + \beta^t\mathbf{X}^t\mathbf{W}\mathbf{X}\beta$$

Since $\mathbf{W}^t = \mathbf{W}$ (as it is a diagonal matrix) and the middle two terms are scalars which are transposes of each other ($\mathbf{Y}^t\mathbf{W}\mathbf{X}\beta = (\mathbf{Y}^t\mathbf{W}\mathbf{X}\beta)^t = \beta^t\mathbf{X}^t\mathbf{W}\mathbf{Y}$), the function simplifies to:

$$L(\beta) = \mathbf{Y}^t\mathbf{W}\mathbf{Y} - 2\mathbf{X}^t\mathbf{W}\mathbf{Y}\beta + \beta^t\mathbf{X}^t\mathbf{W}\mathbf{X}\beta$$

Taking the derivative with respect to β and setting it to zero (the first-order condition):

$$\frac{\partial L}{\partial \beta} = -2\mathbf{X}^t\mathbf{W}\mathbf{Y} + 2\mathbf{X}^t\mathbf{W}\mathbf{X}\beta = \mathbf{0}$$

Solving for $\hat{\beta}$:

$$\hat{\beta} = (\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W}\mathbf{Y}$$

The second derivative is $\frac{\partial^2 L}{\partial \beta^2} = 2\mathbf{X}^t\mathbf{W}\mathbf{X}$, which is positive definite (since $\mathbf{W}$ is positive definite due to the kernel function), confirming that $\hat{\beta}$ is a minimum.

The estimator $\hat{m}(x)$ is the first component of $\hat{\beta}$, so $\hat{m}(x) = \hat{\beta}_0(x)$. Denoting the first row of $(\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W}$ as $(w_{n,1}, \dots, w_{n,n})$, we have:

$$\hat{\beta}_0 = \sum_{i=1}^{n} w_{n,i} Y_i$$

Thus, the Local Linear Regression is a Linear Smoother.

2) Derivation of the Explicit Expression for the Weights

As demonstrated:

$$\begin{pmatrix} \hat{\beta}_0(x) \\ \hat{\beta}_1(x) \end{pmatrix} = (\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W}\mathbf{Y}$$

We compute $\mathbf{X}^t\mathbf{W}\mathbf{X}$:

$$\mathbf{X}^t\mathbf{W}\mathbf{X} = \begin{pmatrix} 1 & \dots & 1 \\ X_1 - x & \dots & X_n - x \end{pmatrix} \mathbf{W} \begin{pmatrix} 1 & X_1 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{pmatrix}$$

Multiplying the first two matrices ($\mathbf{X}^t\mathbf{W}$):

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

Carrying out the final matrix multiplication:

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

Using the notation $S_{n,k} = \frac{1}{nh}\sum_{i=1}^{n}(X_i - x)^k K\left(\frac{X_i - x}{h}\right)$ we have:

$$\mathbf{X}^t\mathbf{W}\mathbf{X} = nh\begin{pmatrix} S_{n,0} & S_{n,1} \\ S_{n,1} & S_{n,2}\end{pmatrix}$$

The inverse of $\mathbf{X}^t\mathbf{W}\mathbf{X}$ is:

$$(\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1} = \frac{1}{nh}\begin{pmatrix} S_{n,0} & S_{n,1} \\ S_{n,1} & S_{n,2}\end{pmatrix}^{-1} = \frac{1}{nh(S_{n,0}S_{n,2} - S_{n,1}^2)}\begin{pmatrix} S_{n,2} & -S_{n,1} \\ -S_{n,1} & S_{n,0}\end{pmatrix}$$

Therfore,

$$\begin{aligned}(\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W} = &\frac{1}{nh(S_{n,0}S_{n,2} - S_{n,1}^2)}\begin{pmatrix} S_{n,2} & -S_{n,1} \\ -S_{n,1} & S_{n,0}\end{pmatrix} \\ &\cdot \begin{pmatrix} K\left(\frac{X_1 - x}{h}\right) & \dots & K\left(\frac{X_n - x}{h}\right) \\ (X_1 - x)K\left(\frac{X_1 - x}{h}\right) & \dots & (X_n - x)K\left(\frac{X_n - x}{h}\right)\end{pmatrix}\end{aligned}$$

The estimator $\hat{\beta}_0(x)$ is found by taking the dot product of the first row of the matrix $(\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W}$ with the vector $\mathbf{Y}$. Hence, the weights $w_{n,i}(x)$ are the entries of the first row of $(\mathbf{X}^t\mathbf{W}\mathbf{X})^{-1}\mathbf{X}^t\mathbf{W}$:

$$\begin{aligned} w_{n,i}(x) &= \frac{1}{nh(S_{n,0}S_{n,2} - S_{n,1}^2)}\left[S_{n,2}K\left(\frac{X_i - x}{h}\right) - S_{n,1}(X_i - x)K\left(\frac{X_i - x}{h}\right)\right] \\ &= \frac{S_{n,2} - S_{n,1}(X_i - x)}{nh(S_{n,0}S_{n,2} - S_{n,1}^2)}K\left(\frac{X_i - x}{h}\right)\end{aligned}$$

3) Proof that the Weights Satisfy $\sum_{i=1}^{n} w_{n,i}(x) = 1$

The sum of the weights is:

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

Factoring out the common terms:

$$= \frac{1}{nh(S_{n,0}S_{n,2} - S_{n,1}^2)}\sum_{i=1}^{n}\left[S_{n,2}K\left(\frac{X_i - x}{h}\right) - S_{n,1}(X_i - x)K\left(\frac{X_i - x}{h}\right)\right]$$

Separating the summation:

$$= \frac{1}{nh(S_{n,0}S_{n,2} - S_{n,1}^2)}\left[S_{n,2}\sum_{i=1}^{n}K\left(\frac{X_i - x}{h}\right) - S_{n,1}\sum_{i=1}^{n}(X_i - x)K\left(\frac{X_i - x}{h}\right)\right]$$

Using the definitions of $S_{n,0}$ and $S_{n,1}$:

$$\sum_{i=1}^{n} K\left(\frac{X_i - x}{h}\right) = nhS_{n,0}$$

$$\sum_{i=1}^{n} (X_i - x)K\left(\frac{X_i - x}{h}\right) = nhS_{n,1}$$

Substituting these back into the expression:

$$\sum_{i=1}^{n} w_{n,i}(x) = \frac{1}{nh(S_{n,0}S_{n,2} - S_{n,1}^2)} \left[S_{n,2}(nhS_{n,0}) - S_{n,1}(nhS_{n,1})\right]$$

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

This proves that the weights satisfy the necessary property: $\sum_{i=1}^{n} w_{n,i}(x) = 1$.

Practical exercise: Local Linear Regression Bandwidth Estimation

The goal of this practical exercise is to study the **Plug-in method** for estimating the asymptotically optimal bandwidth ($\hat{h}_{\text{AMISE}}$) in **Local Linear Regression (LLR)**. This method relies on estimating the residual variance (σ^2) and the second derivative integral ($\theta_{22} = \int (m''(x))^2 f(x)dx$) using a block-wise polynomial fit, with the number of blocks (N) determined by the Mallows' C_p criterion.

Description of the Simulation Study

The estimation procedure $\hat{h}_{\text{AMISE}}$ was tested through three distinct experiments, based on data generated from the model $Y = m(X) + \epsilon$, where $m(x) = \sin(1/(x/3 + 0.1))$ and X follows a Beta distribution. The error term ϵ is $N(0, \sigma^2 = 1^2)$. $R = 50$ repetitions were used for all estimates.

Plug-in Method and C_p Criterion

The estimated optimal bandwidth is given by the formula for the quartic kernel:

$$\hat{h}_{\text{AMISE}} = n^{-1/5} \left(\frac{35\hat{\sigma}^2}{\hat{\theta}_2} \right)^{1/5},$$

where $\hat{\sigma}^2$ and $\hat{\theta}_2$ are estimated using a piecewise 4th-degree polynomial fit over N blocks. The optimal number of blocks N_{opt} is chosen by minimizing Mallows' C_p:

$$C_p(N) = \frac{RSS(N)}{\frac{RSS(N_{\max})}{(n - 5N_{\max})}} - (n - 10N),$$

where $RSS(N) = \sum_{i=1}^{n} \sum_{j=1}^{N} \left\{ Y_i - \hat{m}^j(X_i) \right\}^2 \mathbf{1}_{X_i \in X_j}$, and

$$N_{\max} = \max \left\{ \min \left(\left\lfloor \frac{n}{20} \right\rfloor, 5 \right), 1 \right\}.$$

3.1 Impact of the Number of Blocks (N)

This experiment investigates the sensitivity of the C_p criterion and $\hat{h}_{\text{AMISE}}$ to the choice of the number of blocks N. A large sample size of $n = 2000$ was used with $X \sim \text{Beta}(1, 1)$ (Uniform) to stabilize the estimates.

Analysis of Results

Figure 1 shows the trend of the average $\hat{h}_{\text{AMISE}}$ as N increases, while Figure 2 displays the corresponding Mallows' C_p values.

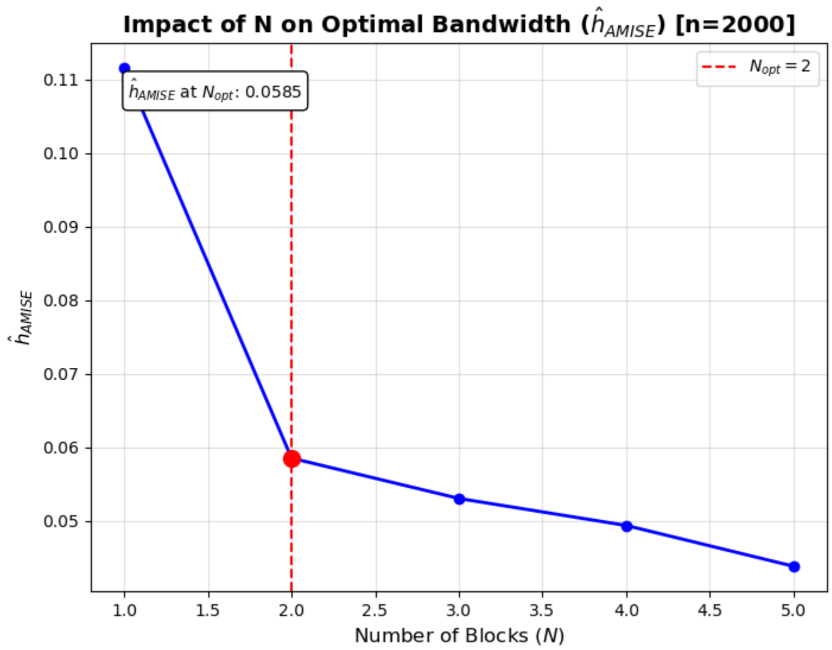

Figure 1: Mean $\hat{h}_{\text{AMISE}}$ as a function of the number of blocks N ($n = 2000$).

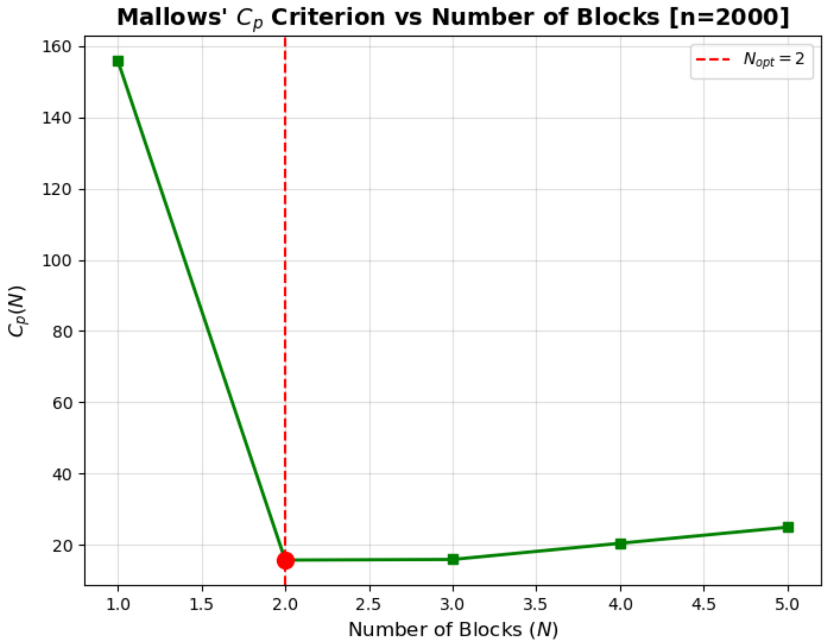

Figure 2: Mean Mallows' $C_p(N)$ criterion as a function of the number of blocks N ($n = 2000$).

The analysis reveals that:

- The minimum of the $C_p(N)$ curve occurs at $N_{\text{opt}} = 2$.
- The optimal bandwidth estimate chosen by the C_p criterion is $\hat{h}_{\text{AMISE}}(N_{\text{opt}} = 2) \approx$ **0.0585**.

It is noted that $\hat{h}_{\text{AMISE}}$ consistently **decreases as** N increases. This is expected because a higher number of blocks (N) allows the piecewise polynomial model to better capture the function's high-frequency curvature (higher $\hat{\theta}_2$), thus requiring a smaller bandwidth ($\hat{h}_{\text{AMISE}} \propto 1/\hat{\theta}_2^{1/5}$). However, the C_p criterion selects $N_{\text{opt}} = 2$ as the optimal complexity, balancing the reduction in bias (lower $\hat{h}$) against the increased variance associated with having too many blocks.

3.2 Impact of the Sample Size (n)

This experiment examines how $\hat{h}_{\text{AMISE}}$ scales with the sample size n, using $X \sim \text{Beta}(1, 1)$ and fixing $N = N_{\text{opt}}$ (implicitly, by using the C_p selection). The theoretical scaling for the LLR bandwidth is $h \propto n^{-1/5}$, implying a slope of -0.200 in the $\log(h)$ vs $\log(n)$ plot.

Analysis of Results

Figure 3 shows the log-log plot of the mean $\hat{h}_{\text{AMISE}}$ against the sample size n.

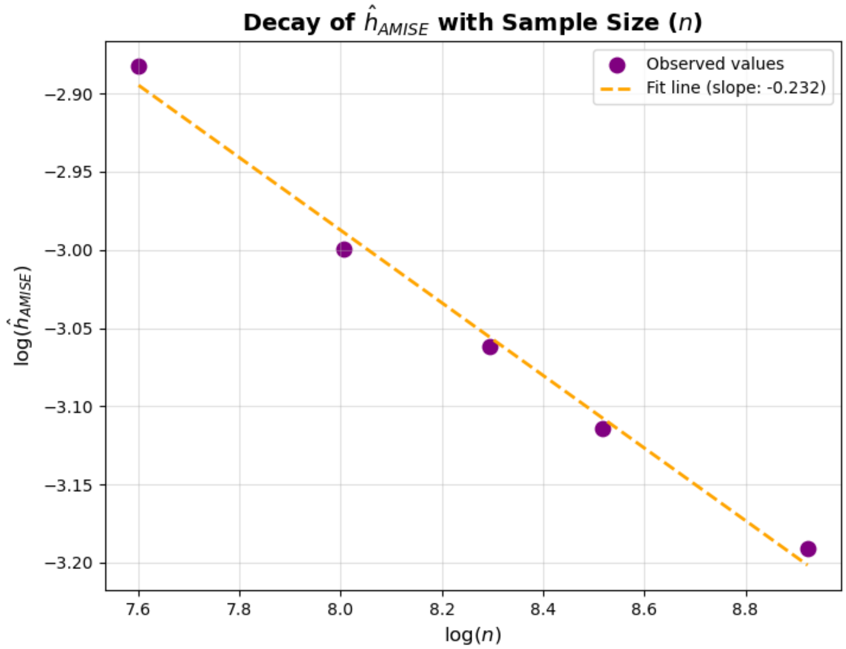

Figure 3: Log-log plot of $\hat{h}_{\text{AMISE}}$ vs sample size n. The slope represents the decay rate.

- The regression line fitted to the observed data points yields an estimated slope of approximately $\mathbf{-0.232}$.
- This estimated slope is extremely close to the theoretical value of -0.200 predicted by the $\mathcal{O}(n^{-1/5})$ rate of convergence for the optimal bandwidth of Local Linear Regression.
- This validates the entire $\hat{h}_{\text{AMISE}}$ estimation procedure, demonstrating that it correctly captures the asymptotic decay rate with respect to the sample size n.

3.3 Impact of Covariate Density Shape

In this final experiment, we fix the sample size at $n = 2000$ and use a fixed number of blocks $N = 5$. We investigate the influence of the covariate density $f(x)$, by sampling X from five different Beta distributions.

Analysis of Results

Figure 4 compares the mean $\hat{h}_{\text{AMISE}}$ estimates across the different density shapes.

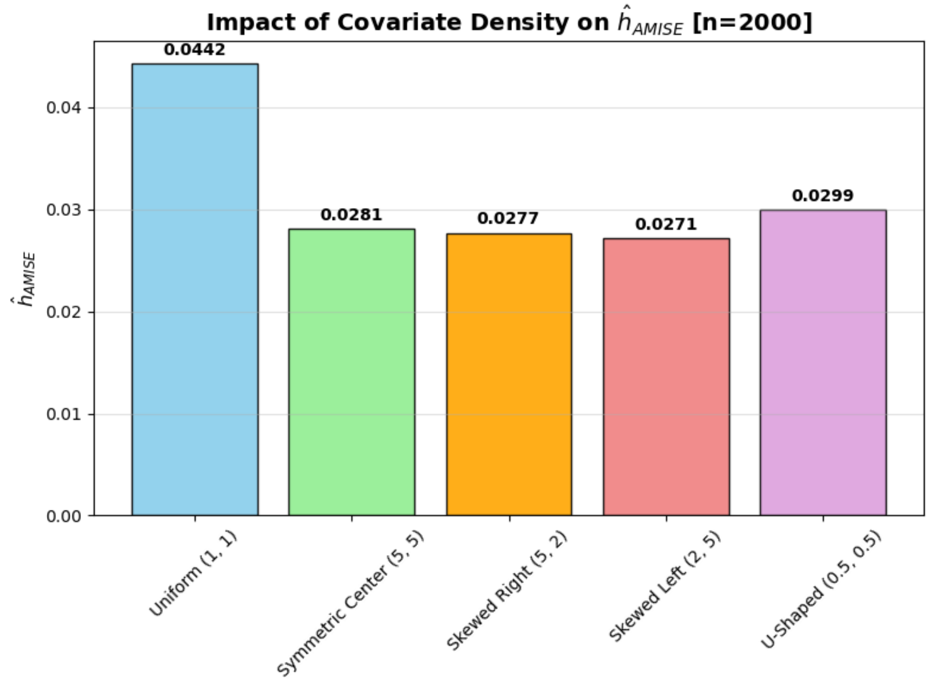

Figure 4: Mean $\hat{h}_{\text{AMISE}}$ for different Beta covariate densities ($n = 2000$, $N = 5$).

The results clearly show that the covariate density significantly influences the optimal bandwidth:

- **Largest $\hat{h}_{\text{AMISE}}$ (Least Smoothing):** The **Uniform (Beta$(1, 1)$)** distribution yields the highest bandwidth (**0.0442**). This reflects its low overall curvature variation.
- **Smallest $\hat{h}_{\text{AMISE}}$ (Most Smoothing):** The **Skewed Left (Beta$(2, 5)$)** distribution results in the lowest bandwidth (**0.0271**). This is driven by the severe sparsity left in the critical high-curvature region near $x = 0$.
- **Non-Uniform Cluster:** All non-uniform distributions (Skewed Left, Skewed Right, Symmetric Center, U-Shaped) are tightly clustered between **0.0271** and **0.0299**, confirming that non-uniform data density generally requires **more aggressive smoothing** (smaller $\hat{h}_{\text{AMISE}}$) compared to the Uniform case.