Chapter 1. Nonparametric Curve Estimation Part 2

January 20, 2007

1 Estimation of density function

Suppose $X_1, \cdots, X_n$ are IID and have a common density function $f(x)$. We can draw the histogram to show the density function.

Example 1.1 *Suppose we draw 500 observations from N(0, 1). then its histogram is as follows (left panel). For the data in* **data** *the histogram is the right panel*

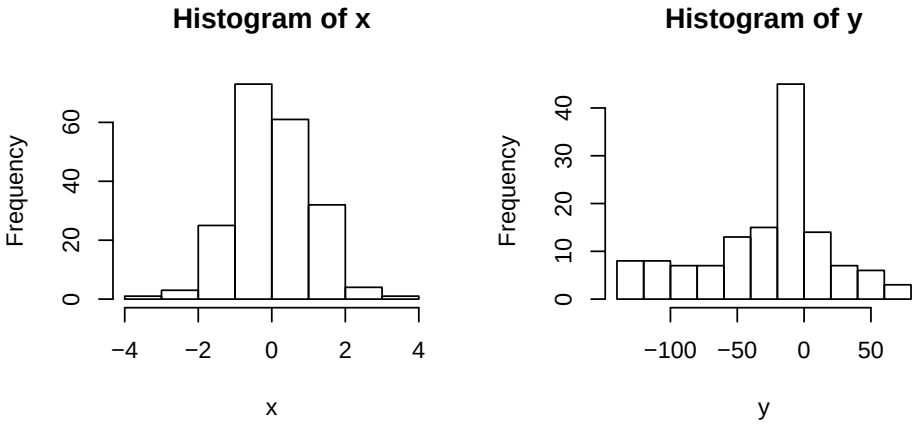

Figure 1: calculation for example 1.1: blue curves are true density function and the red lines are the estimated density functions. **(code)**

Suppose $K()$ is a kernel function (a symmetric density function). We estimate the density by

$$\hat{f}(x) = (nh)^{-1} \sum_{i=1}^{n} K(\frac{x - X_i}{h})$$

Then it is easy to see that $\hat{f}(x)$ is density function because

$$\int \hat{f}(x)dx = 1.$$

It is more convenient to write it as

$$\hat{f}(x) = (n)^{-1} \sum_{i=1}^{n} K_h(x - X_i)$$

where

$$K_h(x - X_i) = \frac{1}{h} K(\frac{x - X_i}{h}).$$

Lemma 1.2 *Suppose $X_1, \cdots, X_n$ are IID and have a common density function $f(x)$.*

$$\hat{f}(x) \quad \rightarrow \quad f(x), \qquad as\ n \rightarrow \infty\ and h \rightarrow 0.$$

In other words, $\hat{f}(x) \stackrel{def}{=} n^{-1} \sum_{i=1}^{n} K_h(X_i - x)$ is a consistent estimator of $f(x)$.

In practice, the bandwidth is chosen by the rule-of-thumb bandwidth selection as

$$h = 1.06 s_x n^{-1/5}.$$

where

$$s_x = \sqrt{n^{-1} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Example 1.3 (Simulation) *samples $X_1, \cdots, X_n$ with size n are drawn from population $X \sim N(0, 1)$. We can estimate the density function of X as*

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{X_i - x}{h})$$

where h is the bandwidth and $K(x)$ is a symmetric density function. h is selected by the rule-of-thumb. The estimated density is shown in Fig. 2

Example 1.4 *The motorcycle data set: n = 133, The estimated pdf of X*

$$\hat{f}_n(x) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{X_i - x}{h})$$

where h is the bandwidth and $K(x)$ is a symmetric density function. h is selected by the rule-of-thumb

The estimated pdf of Y

$$\hat{f}_n(y) = \frac{1}{nh} \sum_{i=1}^{n} K(\frac{Y_i - y}{h})$$

where h is the bandwidth and $K(x)$ is a symmetric density function. h is selected by the rule-of-thumb. The estiamted density is shown in Fig. 3

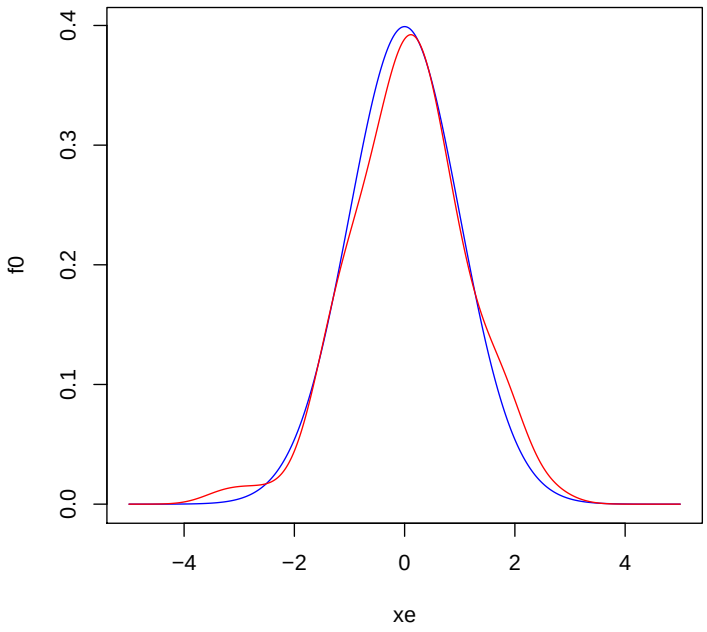

Figure 2: calculation for Example 1.3: blue curves are true density function and the red lines are the estimated density functions. **(code)**

2 Statistical properties of the kernel smoothing

We need to check the errors of the estimators.

2.1 NW-estimator

Consider model

$$Y = m(X) + \varepsilon$$

where ε is independent of X. For easy of exposition, we assume that $m(x)$ has third order derivative. Again, suppose we have the random observations

$$\begin{array}{ll} \text{obs. 1} & X_1 \quad Y_1 \\ \text{obs. 2} & X_2 \quad Y_2 \\ & \cdots \\ \text{obs. n} & X_n \quad Y_n \end{array}$$

or we can write the model as

$$Y_i = m(X_i) + \varepsilon_i$$

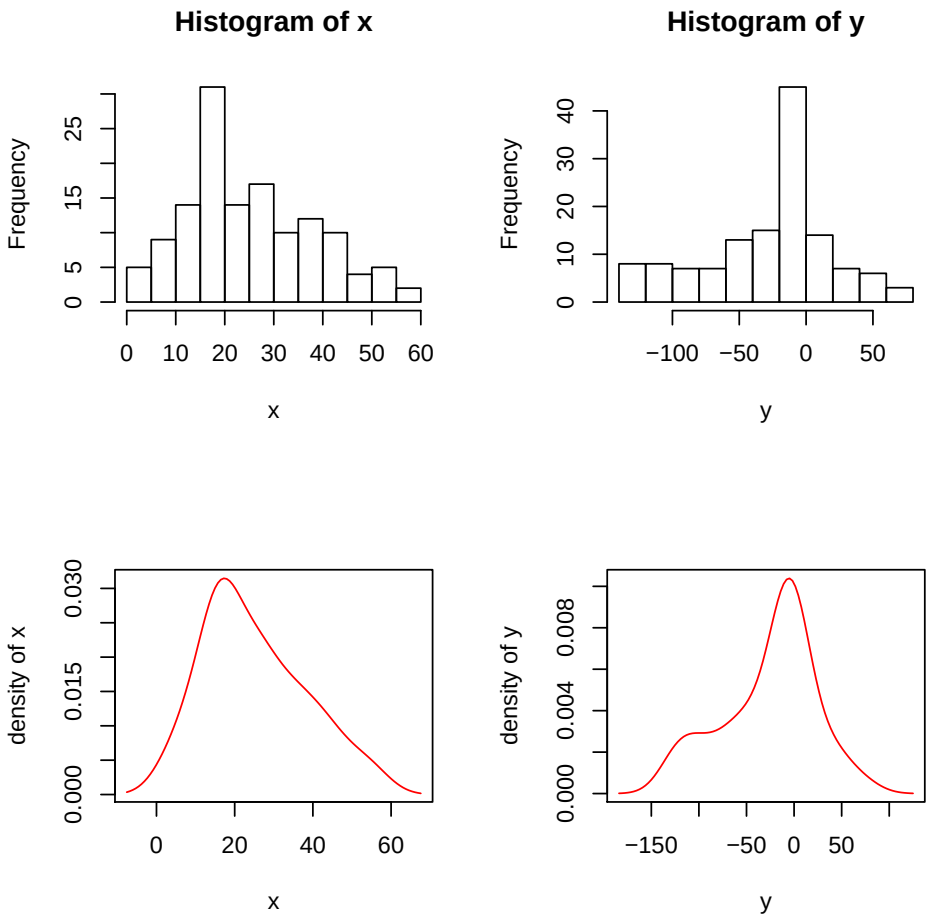

Figure 3: Calculation of Example 1.4. the estimated density functions. **(code)**

A. Taylor Expansion of a smoothing function

For any point x, if X_i is close to x we have the Taylor expansion

$$m(X_i) \approx m(x) + m'(x)(X_i - x) + \frac{1}{2}m''(x)(X_i - x)^2$$

The model can be written as

$$Y_i \approx m(x) + m'(x)(X_i - x) + \frac{1}{2}m''(x)(X_i - x)^2\} + \varepsilon_i \tag{2.1}$$

B. expansion of NW-estimator

Recall the NW-estimator is

$$\hat{m}(x) = \sum_{i=1}^{n} K_h(X_i - x)Y_i / \sum_{i=1}^{n} K_h(X_i - x).$$

By (2.1), we have

$$\hat{m}(x) \approx m(x) + \frac{\sum_{i=1}^n K_h(X_i - x)\left[m'(x)(X_i - x) + \frac{1}{2}m''(x)(X_i - x)^2 + \varepsilon_i\right]}{\sum_{i=1}^n K_h(X_i - x)}$$

We have

$$bias(\hat{m}(x)) \approx \frac{\sum_{i=1}^n K_h(X_i - x)\left[m'(x)(X_i - x) + \frac{1}{2}m''(x)(X_i - x)^2\right]}{\sum_{i=1}^n K_h(X_i - x)}$$

and

$$Var(\hat{m}(x)) = Var\Big(\frac{\sum_{i=1}^n K_h(X_i - x)\varepsilon_i}{\sum_{i=1}^n K_h(X_i - x)}\Big)$$

Proposition 2.1 *Suppose that $h \to 0$ and $nh \to \infty$ as $n \to \infty$. Then at every point of continuity of $m(x)$, $f(x)$ with $f(x) > 0$, we have*

$$E|\hat{m}(x) - m(x)|^2 = bias^2 + variance \to 0$$

as $n \to \infty$ and $h \to 0$

Theorem 2.2 *Suppose that $h \to 0$ and $nh \to \infty$ as $n \to \infty$. $EY^2 < \infty$. Then at every point of continuity of $m(x)$, $f(x)$ with $f(x) > 0$, we have*

$$biase \stackrel{def}{=} E\hat{m}(x) - m(x) = c_2\{\frac{1}{2}m''(x) + f^{-1}(x)m'(x)f'(x)\}h^2$$

and

$$variance \stackrel{def}{=} var(\hat{m}(x) - m(x)) \to \frac{d_0\sigma^2}{nhf(x)}.$$

where $d_0 = \int K(v)^2 dv$.

Therefore, $\hat{m}(x)$ is a biased estimator of $m(x)$. The choice of h should minimize

$$\begin{aligned} E|\hat{m}(x) - m(x)|^2 &= biase^2 + variance \\ &= c_2^2\{\frac{1}{2}m''(x) + f^{-1}(x)m'(x)f'(x)\}^2h^4 + \frac{d_0\sigma^2}{nhf(x)} \end{aligned}$$

The optimal bandwidth is than

$$h_{opt} = \Big\{\frac{d_0\sigma^2}{4f(x)c_2^2\{\frac{1}{2}m''(x) + f^{-1}(x)m'(x)f'(x)\}^2}\Big\}^{1/5} n^{-1/5}.$$

Example 2.3 *n samples are drawn from*

$$Y = sin(2\pi X) + 0.2\varepsilon.$$

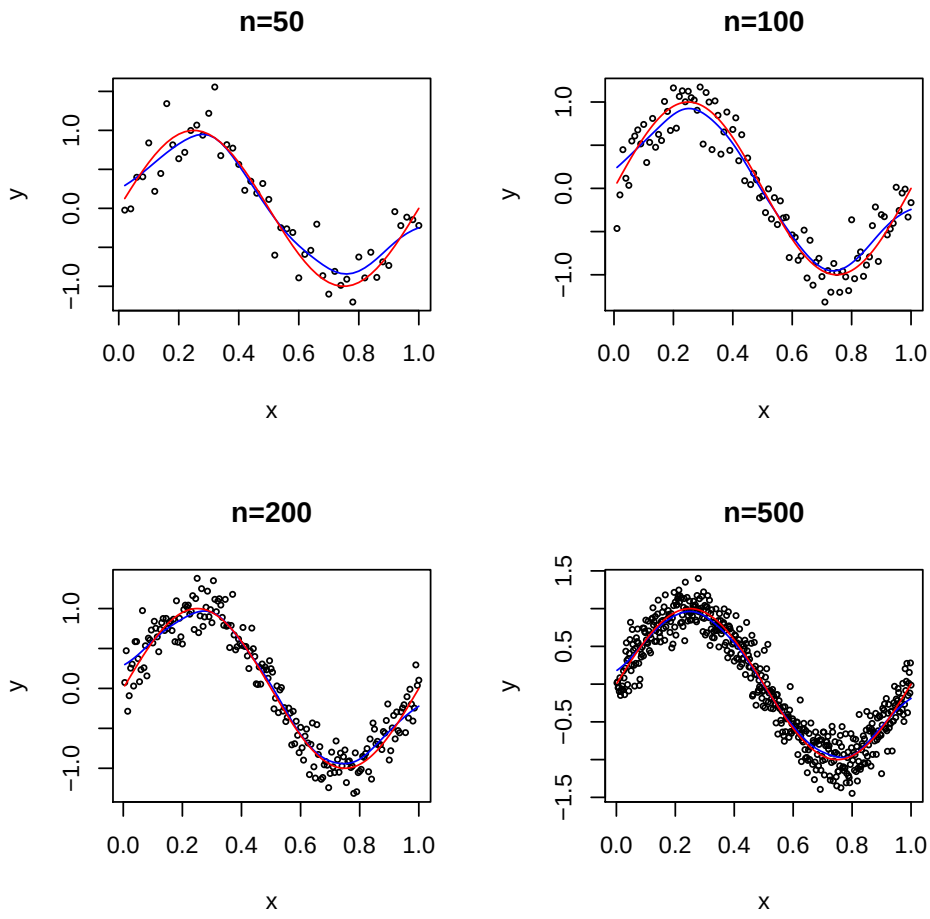

Figure 4: Calculation for Example 2.3 blue curves are true regression function and the red lines are the estimated functions

where $X \in Uniform(0,1)$ *and* $\varepsilon \sim N(0,1)$.

Estimated function is shown in Fig 4

We can see from the figure, as n increase, the estimated curve tends to the true curve funciton. Note that

$$[\sin(2\pi x)]'' = -(2\pi)^2 \sin(2\pi x)$$

which has big absolute value at x = 1/4 and 3/4. Therefore, the estimated curve has big bias at those points.

Example 2.4 *n samples are drawn from*

$$Y = exp(-20X^2) + 0.2\varepsilon.$$

where $X \in Uniform(-1,1)$ *and* $\varepsilon \sim N(0,1)$. *Estimated function is shown in Fig 5*

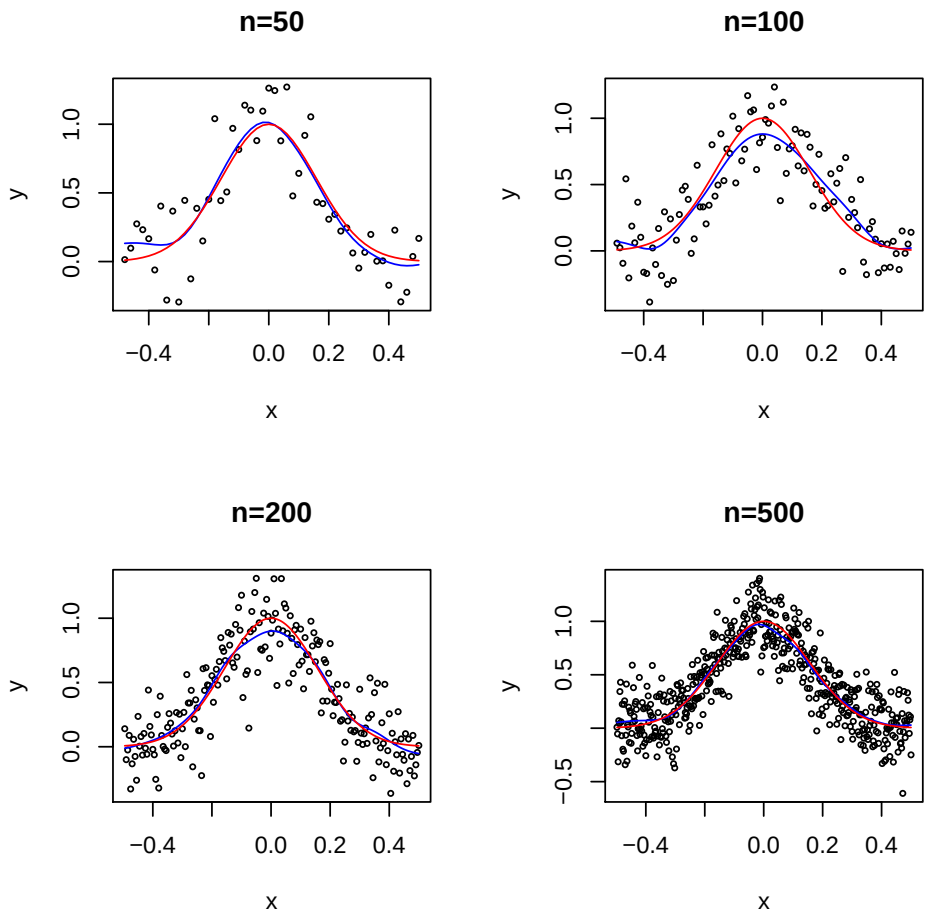

Figure 5: Calculation for Example 2.4: blue curves are true regression function and the red lines are the estimated functions **(code)**

Lemma 2.5 *If $X_1, \cdots, X_n$ are IID and have common density function $f(x)$. If ε is independent of X and $Var(\varepsilon) = \sigma^2$, then*

$$\sqrt{nh} * n^{-1} \sum_{i=1}^{n} K_h(X_i - x)\varepsilon_i \xrightarrow{D} N(0, d_0 f(x)\sigma^2)$$

where $d_0 = \int K(v)^2 dv$.

Theorem 2.6

$$\sqrt{nh}\{\hat{m}(x) - m(x) - \frac{1}{2}c_2[m''(x) + 2f^{-1}(x)m'(x)f'(x)]h^2\} \xrightarrow{D} N(0, \frac{d_0\sigma^2}{f(x)})$$

If $\sqrt{nh}h^2 \to 0$, then

$$\sqrt{nh}\{\hat{m}(x) - m(x)\} \xrightarrow{D} N(0, \frac{d_0\sigma^2}{f(x)})$$

The point-wise confidence band is

$$P\Big\{\hat{m}(x) - 1.96\sqrt{\frac{d_0\sigma^2}{nhf(x)}} \le m(x) \le \hat{m}(x) + 1.96\sqrt{\frac{d_0\sigma^2}{nhf(x)}}\Big\} \approx 0.95$$

In using the above theorem to draw the confidence band, we need to calculate a number of values

1. Gaussian Kernel:

$$c_0 = 1, c_1 = 0, c_2 = 1, d_0 = 0.2821$$

For Epanechnikov kernel

$$c_0 = 1, c_1 = 0, c_2 = 0.2, d_0 = 0.6$$

2. σ^2

$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \hat{m}(X_i))^2$$

3. the density function of X

$$\hat{f}_n(x) = n^{-1}\sum_{i=1}^{n} K_h(X_i - x)$$

Example 2.7 (Simulation) *consider model* $Y = \cos(\pi X) + 0.2\varepsilon$, *where* $X \sim uniform(0,1)$ *and* $\varepsilon \sim N(0,1)$. *with sample size* n. *Consider* $h = 0.1$. *The estimated regression curve*

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K_h(X_i - x)Y_i}{\sum_{i=1}^{n} K_h(X_i - x)}$$

The estimated pdf of X.

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{X_i - x}{h})$$

where h *is the bandwidth and* $K(x)$ *is a symmetric density function. The esimator of* σ^2 *is*

$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \hat{m}(X_i))^2.$$

The estimated function is shown in Fig 6.

Example 2.8 *The motorcycle data set:* $n = 133$, *consider* $h = 1.5$. *The estimated regression curve*

$$\hat{m}(x) = \frac{\sum_{i=1}^{n} K_h(X_i - x)Y_i}{\sum_{i=1}^{n} K_h(X_i - x)}$$

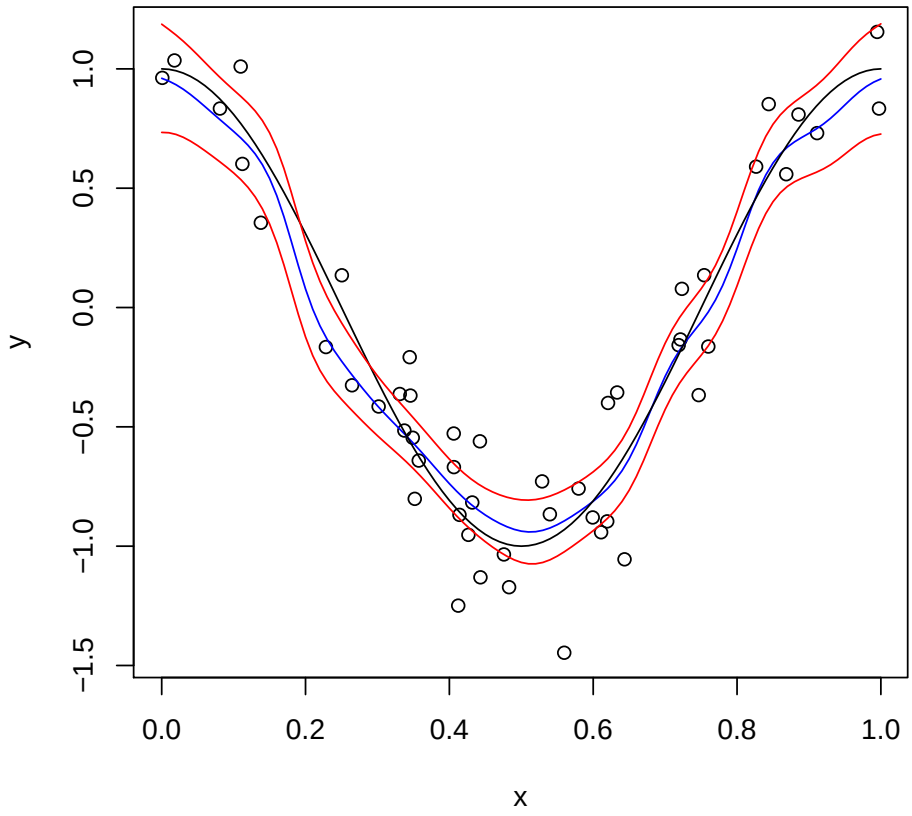

Figure 6: calculation for Example 2.7 The black line is the true function. The blue line in the central is the estimated regression function, the upper and lower red lines are the 95% point-wise confidence band. **(code)**

The estimated pdf of X.

$$\hat{f}_n(x) = \frac{1}{nh}\sum_{i=1}^{n} K(\frac{X_i - x}{h})$$

where h is the bandwidth and $K(x)$ is a symmetric density function. The esimator of σ^2 is

$$\hat{\sigma}^2 = n^{-1}\sum_{i=1}^{n}(Y_i - \hat{m}(X_i))^2.$$

The estimated function is shown in Fig 7

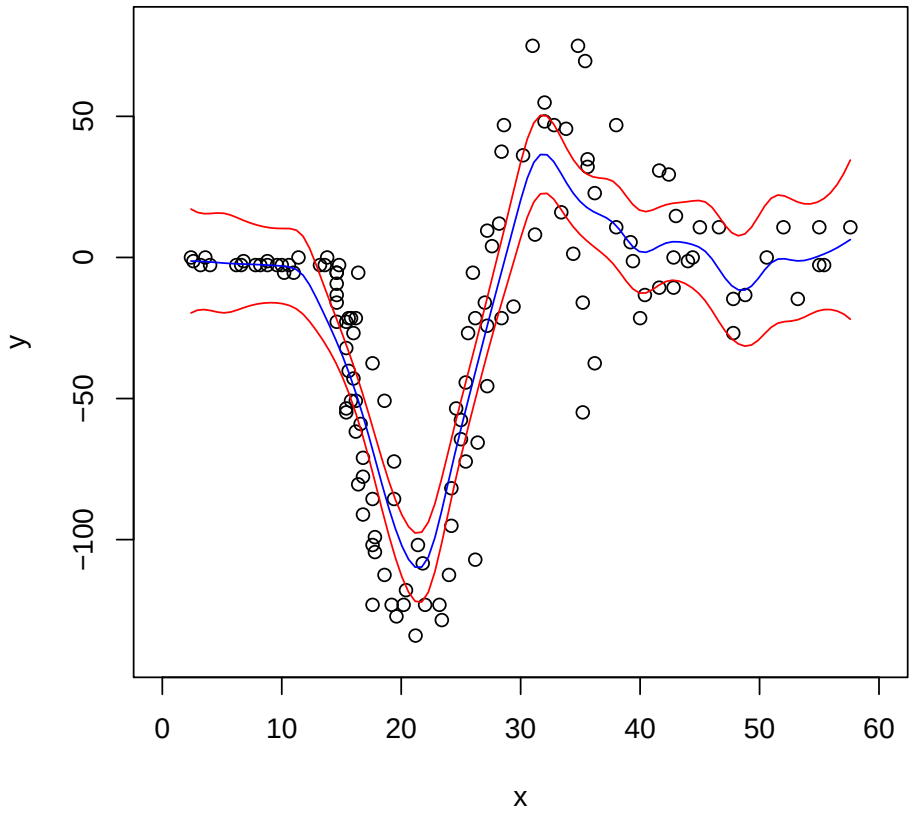

Figure 7: Calculation for Example 2.8. The line in the central is the estimated regression function, the upper and lower lines are the 95% point-wise confidence band. **(code)**