The Nadaraya-Watson Estimator - Continued

Example

As an example, we will consider $n = 100$ bivariate observations $(Y_i, x_i), i = 1, \ldots, n$ from the model

$$Y_i = m(x_i) + \epsilon_i, \quad i = 1, \ldots, n$$

where the regression function $m(x) = \sin(2 * \pi * x^3)^3$, the x_i's are iid $U(0,1)$ random variables and the errors ϵ_i are iid $0.2 * t(15)$ random variables. The model used here to simulate the data is similar to the one used earlier in the module, except now the x-variables are random $U(0,1)$ observations, rather than being equally spaced in $(0,1)$.

The method of least squares cross-validation (lscv) was used to select an appropriate degree of smoothing. The search was narrowed down to the interval $(0, 0.1)$ and a plot of the function $CV(h)$ is plotted in figure 1 calculated for 25 equi-spaced values of h in this interval.

The minimizing value of h was found to be $h = 0.034$. A scatterplot of the data with the smooth based on the lscv-optimal h and also the true function are shown in figure 2.

The estimate is a very good fit on the interval $0.4, 1.0)$ but is a bit noisy on $(0, 0.4)$ where the true curve is quite flat.

Exact bias, variance and mse

We can evaluate the exact expectation and variance of the Nadaraya-Watson estimator of $\hat{m}(x)$ as follows:

$$\begin{aligned} E(\hat{m}(x)) &= \frac{\sum_{i=1}^n K_{h_x}(x - x_i) m(x_i)}{\sum_{i=1}^n K_{h_x}(x - x_i)} \\ &= \sum_{i=1}^n W_{h_x}(x, x_i) m(x_i) \end{aligned}$$

which is a smooth of the true regression function values $m(x_i)$. The variance is given by

$$\begin{aligned} V(\hat{m}(x)) &= \sigma_\epsilon^2 \frac{\sum_{i=1}^n K_{h_x}(x - x_i)^2}{(\sum_{i=1}^n K_{h_x}(x - x_i))^2} \\ &= \sigma_\epsilon^2 \sum_{i=1}^n W_{h_x}(x, x_i)^2 \end{aligned}$$

Figure 3 shows a plot of the true expected values of the estimator (based on h=0.034) together with the true regression curve. For this estimator and the given h the expected values are very close to the true ones. This is further illustrated in the plot of the bias in figure 4.

The true variances of $\hat{m}(x)$ are plotted in figure 5. This is quite a noisy curve with the largest variances being around the x-values $0.33, 0.69 - 0.74$ and 0.88. The ensuing mse curve in figure 6 is a similar shape to the variance curve as, for these data and the estimator used, the variances are generally much larger than the squared bias's.

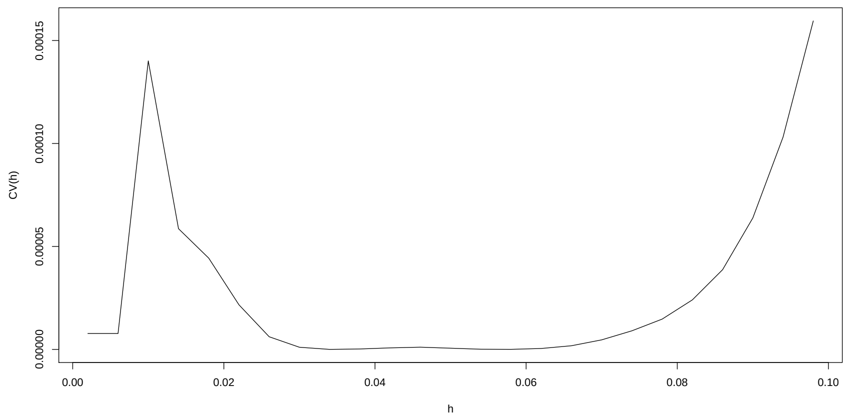

Figure 1: Least-squares cross-validation for the simulated data. A Nadaraya-Watson estimate using a biweight (quartic) kernel was used.

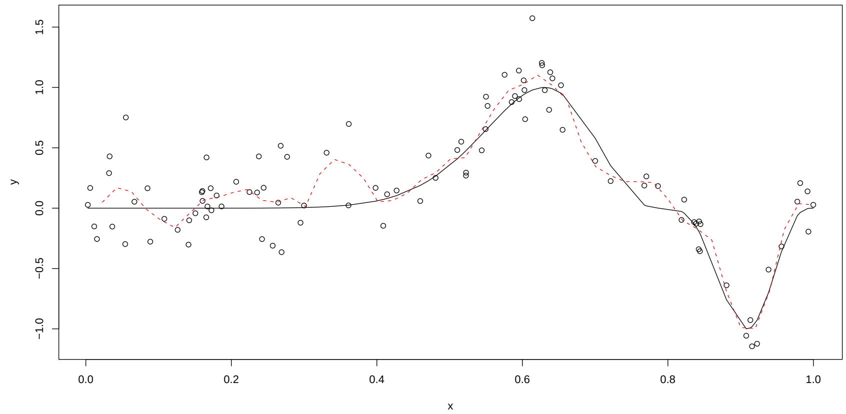

Figure 2: The simulated data with the true regression curve (black line) and Nadaraya-Watson smooth using a biweight kernel and h=0.034 (red line)

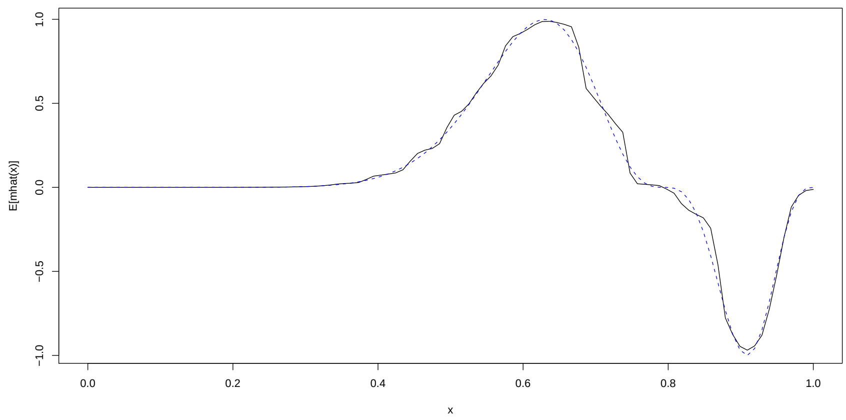

Figure 3: Exact expectation of the Nadaraya-Watson estimator of the simulated regression data based on $h = 0.034$ - black line; true curve - blue dashed line. A biweight (quartic) kernel was used.

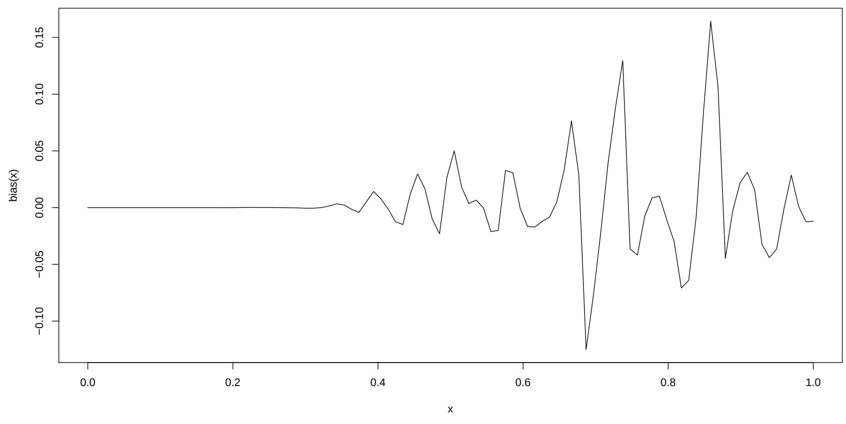

Figure 4: Exact bias of the Nadaraya-Watson estimator of the simulated regression data based on $h = 0.034$. A biweight (quartic) kernel was used.

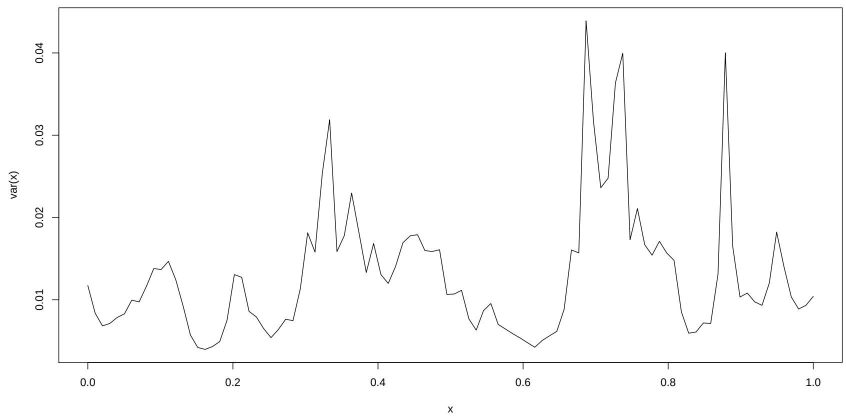

Figure 5: Exact variance of the Nadaraya-Watson estimator of the simulated regression data based on $h = 0.034$. A biweight (quartic) kernel was used.

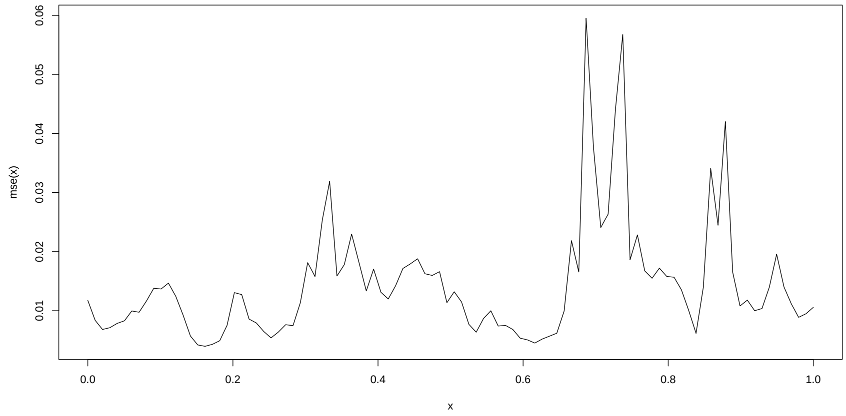

Figure 6: Exact mse of the Nadaraya-Watson estimator of the simulated regression data based on $h = 0.034$. A biweight (quartic) kernel was used.