The Local Polynomial Regression Estimator

Deriving the estimator

Let $(x_1, Y_1), \ldots, (x_n, Y_n)$ be a random sample of bivariate data that we have available to estimate the unknown regression function $m(x) = E[Y \mid X = x]$. Using Taylor Series, we can approximate $m(x)$, where x is close to a point x_0, as follows:

$$\begin{aligned} m(x) &\approx m(x_0) + m^{(1)}(x_0)(x - x_0) + \frac{m^{(2)}(x_0)}{2!}(x - x_0)^2 + \ldots \\ &\quad \ldots + \frac{m^{(p)}(x_0)}{p!}(x - x_0)^p \\ &= m(x_0) + \beta_1(x - x_0) + \beta_2(x - x_0)^2 + \ldots + \beta_p(x - x_0)^p \end{aligned}$$

provided that all the required derivatives exist. This is a polynomial of degree p. We can then use this in a minimization problem with the data on x and Y. This is the local polynomial regression problem in which we use the data to estimate that polynomial of degree p which best approximates $m(x)$ in a small neighborhood around the point x_0. ie. we minimize with respect to $\beta_0, \beta_1, \ldots, \beta_p$ the function

$$\sum_{i=1}^{n} \{Y_i - \beta_0 - \beta_1(x_i - x_0) - \ldots - \beta_p(x_i - x_0)^p\}^2 K\left(\frac{x_i - x_0}{h}\right)$$

This is a weighted least squares problem where the weights are given by the kernel functions $K((x_i - x_0)/h)$.

It is convenient to define the following vectors and matrices:

$$\mathbf{X}_{x_0} = \begin{bmatrix} 1, & (x_1 - x_0), & \ldots & , (x_1 - x_0)^p \\ 1, & (x_2 - x_0), & \ldots & , (x_2 - x_0)^p \\ \vdots & \vdots & \ddots & \vdots \\ 1, & (x_n - x_0), & \ldots & , (x_n - x_0)^p \end{bmatrix}$$

$$\mathbf{Y} = (Y_1, Y_2, \ldots, Y_n)^T$$

$$\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_p)^T$$

$$\mathbf{W}_{x_0} = \begin{bmatrix} K((x_1 - x_0)/h), & 0 & \dots & ,0 \\ 0, & K((x_2 - x_0)/h), & \dots & ,0 \\ \vdots & \vdots & \ddots & \vdots \\ 0, & 0, & \dots & ,K((x_n - x_0)/h) \end{bmatrix}$$

Note that, because the kernel K is symmetric, we could have written the argument of K as $(x_0 - x_i)/h$. However, the notation used here emphasizes the fact that the local polynomial regression is a weighted regression using data centered around x_0. The least squares problem is then to minimize the weighted sum-of-squares function

$$(\mathbf{Y} - \mathbf{X}_{x_0}\boldsymbol{\beta})^T \mathbf{W}_{x_0} (\mathbf{Y} - \mathbf{X}_{x_0}\boldsymbol{\beta})$$

with respect to the parameters $\boldsymbol{\beta}$. The solution is

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}_{x_0}^T \mathbf{W}_{x_0} \mathbf{X}_{x_0})^{-1} (\mathbf{X}_{x_0}^T \mathbf{W}_{x_0} \mathbf{Y})$$

provided $(\mathbf{X}^T\mathbf{W}\mathbf{X})$ is a nonsingular matrix. The quantity $m(x_0)$ is then estimated by the fitted intercept parameter (ie. by $\hat{\beta}_0$) as this defines the position of the estimated local polynomial curve at the point x_0. By varying the value of x_0, we can build up an estimate of the function $m(x)$ over the range of the data. We have

$$\hat{m}(x_0) = \mathbf{e}_1^T (\mathbf{X}_{x_0}^T \mathbf{W}_{x_0} \mathbf{X}_{x_0})^{-1} (\mathbf{X}_{x_0}^T \mathbf{W}_{x_0} \mathbf{Y})$$

where the vector $\mathbf{e}_1$ is of length $p + 1$ and has a 1 in the first position and 0's elsewhere.

When $p = 0$ (local constant), $\hat{m}(x)$ is equivalent to the Nadaraya-Watson estimator

$$\begin{aligned} \hat{m}(x_0) &= \frac{\sum_{i=1}^n K((x_i - x_0)/h) Y_i}{\sum_{i=1}^n K((x_i - x_0)/h)} \\ &= \frac{\sum_{i=1}^n K((x_0 - x_i)/h) Y_i}{\sum_{i=1}^n K((x_0 - x_i)/h)} \\ &= \sum_{i=1}^n W_{h_x}(x_0, x_i) Y_i \end{aligned}$$

When $p = 1$ (local linear), we can express the estimator as

$$\hat{m}(x_0) = \frac{1}{n} \sum_{i=1}^n \frac{(\hat{s}_2(x_0; h) - \hat{s}_1(x_0; h)(x_i - x_0)) K((x_i - x_0)/h) Y_i}{\hat{s}_2(x_0; h)\hat{s}_0(x_0; h) - \hat{s}_1(x_0; h)^2}$$

where $\hat{s}_j(x_0; h) = n^{-1} \sum_{i=1}^n (x_i - x_0)^j K((x_i - x_0)/h)$ for $j = 0, 1, 2$. This is a linear function of the Y_i's and so is a "linear smoother" in the sense that we have previously defined.

Asymptotic MSE properties - the local linear case (p=1)

We will firstly consider the fixed equally-spaced design model where the x-variables are assumed to lie in the interval $(0,1)$ so that $x_i = i/n$ for $i = 1,\ldots,n$.

We will need to make the following assumptions:

(i) The function $m^{(2)}(\cdot)$ is continuous on $(0,1)$.

(ii) The kernel K is symmetric and supported on $(-1,1)$. Also, K has a bounded first derivative.

(iii) The bandwidth h is a sequence of values which depend on the sample size n and satisfying $h \to 0$ and $nh \to \infty$ as $n \to \infty$.

(iv) The point x at which the estimation is taking place satisfies $h < x < 1-h \ \ \forall n \geq n_0$, where n_0 is fixed.

It follows from the definition of the estimator that

$$E[\hat{m}(x)] \ = \ \mathbf{e}_1^T(\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{X}_x^T\mathbf{W}_x\mathbf{M}$$

where the vector $\mathbf{M} = (m(x_1),\ldots,m(x_n))^T$ contains the true regression function values at each of the x_i's. Note that, in this section we will denote the point at which m is being estimated simply by x rather than x_0, as was done in the first section above.

For local linear regression (the case $p=1$) we have that

$$\mathbf{X}_x = \begin{bmatrix} 1, & (x_1 - x) \\ 1, & (x_2 - x) \\ \vdots & \vdots \\ 1, & (x_n - x) \end{bmatrix}$$

By Taylor's theorem, for any $x \in (0,1)$ we can write

$$m(x_i) = m(x) + (x_i - x)m^{(1)}(x) + \frac{1}{2}(x_i - x)^2 m^{(2)}(x) + \ldots$$

so that

$$\mathbf{M} = \mathbf{X}_x \begin{bmatrix} m(x) \\ m^{(1)}(x) \end{bmatrix} + \frac{1}{2}m^{(2)}(x) \begin{bmatrix} (x_1 - x)^2 \\ \vdots \\ (x_n - x)^2 \end{bmatrix} + \ldots$$

The first term in the expansion of $\hat{m}(x)$ is therefore

$$\mathbf{e}_1^T(\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x)^{-1}(\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x) \begin{bmatrix} m(x) \\ m^{(1)}(x) \end{bmatrix} = \mathbf{e}_1^T \begin{bmatrix} m(x) \\ m^{(1)}(x) \end{bmatrix} = m(x)$$

which is the true regression function. The bias of the estimator $\hat{m}(x)$ is then

$$\begin{aligned} E[\hat{m}(x)] - m(x) &= \mathbf{e}_1^T (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} \mathbf{X}_x^T \mathbf{W}_x \left[\frac{1}{2} m^{(2)}(x)\right] \begin{bmatrix} (x_1 - x)^2 \\ \vdots \\ (x_n - x)^2 \end{bmatrix} + \dots \\ &= \mathbf{e}_1^T n (\mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x)^{-1} (n^{-1}) \mathbf{X}_x^T \mathbf{W}_x \left[\frac{1}{2} m^{(2)}(x)\right] \begin{bmatrix} (x_1 - x)^2 \\ \vdots \\ (x_n - x)^2 \end{bmatrix} + \dots \end{aligned}$$

Note that if m is a linear function then $m^{(r)}(x) = 0 \ \ \forall\, r \geq 2$ so that the local linear estimator is exactly unbiased when m is a linear function.

To find the leading bias term for general functions m, note that

$$n^{-1} \mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x = \begin{bmatrix} \hat{s}_0(x;h) & \hat{s}_1(x;h) \\ \hat{s}_1(x;h) & \hat{s}_2(x;h) \end{bmatrix}$$

and

$$n^{-1} \mathbf{X}_x^T \mathbf{W}_x \begin{bmatrix} (x_1 - x)^2 \\ \vdots \\ (x_n - x)^2 \end{bmatrix} = \begin{bmatrix} \hat{s}_2(x;h) \\ \hat{s}_3(x;h) \end{bmatrix}$$

where $\hat{s}_j(x;h) = n^{-1} \sum_{i=1}^n (x_i - x)^j K((x_i - x)/h)$ for $j = 0, 1, 2, 3$.

Since the first derivative $K^{(1)}$ of the kernel is assumed to be bounded, we can approximate the functions $\hat{s}_j(x;h)$ by integrals. In order to do this we need conditions (i)-(iv) above and the sample size n to be sufficiently large. We have

$$\begin{aligned} \hat{s}_j(x;h) &= \int_0^1 (y - x)^j K((y - x)/h) dy + O(n^{-1}) \\ &= h^{j+1} \int_{-x/h}^{(1-x)/h} u^j K(u) du + O(n^{-1}) \\ &= h^{j+1} \int_{-1}^1 u^j K(u) du + O(n^{-1}) \end{aligned}$$

By the symmetry and compact support of K, the odd moments of K are all zero and so we have

$$\begin{aligned} n^{-1} \mathbf{X}_x^T \mathbf{W}_x \mathbf{X}_x &= \begin{bmatrix} \hat{s}_0(x;h) & \hat{s}_1(x;h) \\ \hat{s}_1(x;h) & \hat{s}_2(x;h) \end{bmatrix} \\ &= \begin{bmatrix} h + O(n^{-1}) & O(n^{-1}) \\ O(n^{-1}) & h^3 \sigma_K^2 + O(n^{-1}) \end{bmatrix} \end{aligned}$$

where $\sigma_K^2 = \int_{-1}^{1} u^2 K(u)du$ and

$$\begin{aligned} n^{-1}\mathbf{X}_x^T\mathbf{W}_x \begin{bmatrix} (x_1-x)^2 \\ \vdots \\ (x_n-x)^2 \end{bmatrix} &= \begin{bmatrix} \hat{s}_2(x;h) \\ \hat{s}_3(x;h) \end{bmatrix} \\ &= \begin{bmatrix} h^3\sigma_K^2 + O(n^{-1}) \\ O(n^{-1}) \end{bmatrix} \end{aligned}$$

Some straightforward matrix algebra then leads to the following expression for the leading bias term

$$E\hat{m}(x) - m(x) = \frac{1}{2}h^2\sigma_K^2 m^2(x) + o(h^2) + O(n^{-1})$$

To derive the asymptotic variance of $\hat{m}(x)$ we have

$$\begin{aligned} V(\hat{m}(x)) &= \mathbf{e}_1^T(\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x)^{-1}(\mathbf{X}_x^T\mathbf{W}_x\mathbf{V}\mathbf{W}_x\mathbf{X}_x)(\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{e}_1 \\ &= (n^{-1})\mathbf{e}_1^T n(\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x)^{-1}(n^{-1})(\mathbf{X}_x^T\mathbf{W}_x\mathbf{V}\mathbf{W}_x\mathbf{X}_x)n(\mathbf{X}_x^T\mathbf{W}_x\mathbf{X}_x)^{-1}\mathbf{e}_1 \end{aligned}$$

where $\mathbf{V} = diag(\sigma_\epsilon^2, \ldots, \sigma_\epsilon^2)$. Now, using approximations analogous to those used above we have that

$$\begin{aligned} n^{-1}(\mathbf{X}_x^T\mathbf{W}_x\mathbf{V}\mathbf{W}_x\mathbf{X}_x) &= n^{-1}\sum_{i=1}^{n} K((x_i-x)/h)^2\sigma_K^2 \begin{bmatrix} 1 & x_i-x \\ x_i-x & (x_i-x)^2 \end{bmatrix} \\ &= \begin{bmatrix} h\sigma_K^2 R(K) + o(n^{-1}) & O(n^{-1}) \\ O(n^{-1}) & h^3\sigma_K^2\int u^2K(u)^2du + O(n^{-1}) \end{bmatrix} \end{aligned}$$

where $R(K) = \int K(u)^2du$. These expressions can be combined to obtain

$$V(\hat{m}(x)) = \frac{1}{nh}\sigma_K^2 R(K) + o\{(nh)^{-1}\}$$

In the random design case here, we will assume that the x-variables constitute an independent random sample $(X_1, \ldots, X_n)$ each having the pdf f which has support on $(0, 1)$. We also need to assume that f', the first derivative of f, is continuous.

In the random design setting the calculations of the asymptotic bias and variance of $\hat{m}(x)$ are carried out in an analogous way to those in the fixed design case, provided that we condition on the X_i's. It can be shown that the conditional bias is given by

$$E[\hat{m}(x) - m(x) \mid X_1, \ldots, X_n] = \frac{1}{2}h^2\sigma_K^2 m^{(2)}(x) + o_P(h^2)$$

while the conditional variance is

$$V[\hat{m}(x) \mid X_1, \ldots, X_n] \;=\; \frac{1}{nh}\,\sigma_K^2\,\frac{R(K)}{f(x)} + o_P\{(nh)^{-1}\}$$

Note that all the above results are derived for the case when the weight function is defined as $K((x_i - x)/h)$ but, we could have used the scaled kernel function $(h^{-1})K((x_i - x)/h)$ instead. The solution to the local weighted least squares problem is exactly the same for both weight functions as the latter only scales the former by h^{-1}. The asymptotic biases and variances of the estimators based on these two weight functions are identical but there are some small differences in their derivation when approximating sums of kernel functions by integrals in the functions $\hat{s}_j(x; h)$.

Example

The data used for the example are identical to those used for the example in the Nadaraya-Watson section. ie. we have $n = 100$ 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. We use a local linear regression estimator based on a gaussian kernel with $h = 0.0243$. The function "locpoly" from the R package KernSmooth was used to produce the following plot.

The estimated regression curve in Figure 1 is very similar to that found using the Nadaraya-Watson estimate.

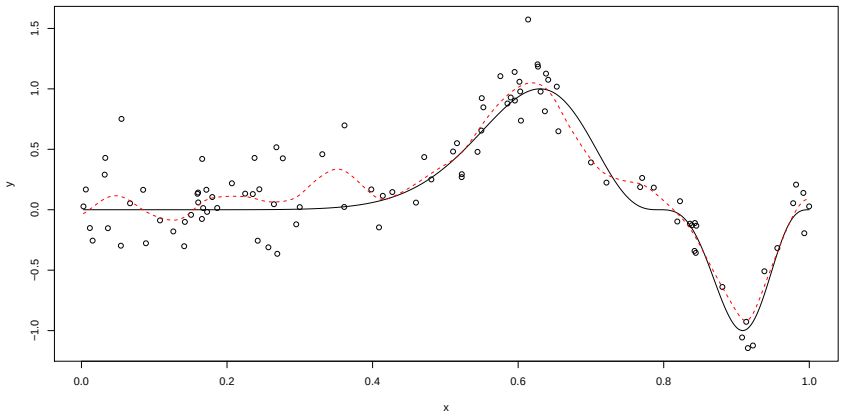

Figure 1: The simulated data with the true regression curve (black line) and local linear smooth using a gaussian kernel and h=0.0243 (red line)