Locally Linear Regression:

There is another local method, locally linear regression, that is thought to be superior to kernel regression. It is based on locally fitting a line rather than a constant. Unlike kernel regression, locally linear estimation would have no bias if the true model were linear. In general, locally linear estimation removes a bias term from the kernel estimator, that makes it have better behavior near the boundary of the $x$’s and smaller MSE everywhere.

To describe this estimator, let $K_h(u) = h^{-r}K(u/h)$ as before. Consider the estimator $\hat{g}(x)$ given by the solution to

$$
\min_{g, \beta} \sum_{i=1}^{n} (Y_i - g - (x - x_i)'\beta)^2 K_h(x - x_i).
$$

That is $\hat{g}(x)$ is the constant term in a weighted least squares regression of $Y_i$ on $(1, x - x_i)$, with weights $K_h(x - x_i)$. For

$$
Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, 
X = \begin{pmatrix} 1 & (x - x_1)' \\ \vdots & \vdots \\ 1 & (x - x_n)' \end{pmatrix}
$$

$$
W = diag(K_h(x - x_1), \ldots, K_h(x - x_n))
$$

and $e_1$ a $(r + 1) \times 1$ vector with 1 in first position and zeros elsewhere, we have

$$
\hat{g}(x) = e'_1 (X'WX)^{-1}X'WY.
$$

This estimator depends on $x$ both through the weights $K_h(x - x_i)$ and through the regressors $x - x_i$.

This estimator is a locally linear fit of the data. It runs a regression with weights that are smaller for observations that are farther from $x$. In contrast, the kernel regression estimator solves this same minimization problem but with $\beta$ constrained to be zero, i.e., kernel regression minimizes

$$
\sum_{i=1}^{n} (Y_i - g)^2 K_h(x - x_i)
$$

Removing the constraint $\beta = 0$ leads to lower bias without increasing variance when $g_0(x)$ is twice differentiable. It is also of interest to note that $\hat{\beta}$ from the above minimization problem estimates the gradient $\partial g_0(x) / \partial x$.

Like kernel regression, this estimator can be interpreted as a weighted average of the $Y_i$ observations, though the weights are a bit more complicated. Let

$$
S_0 = \sum_{i=1}^{n} K_h(x - x_i), 
S_1 = \sum_{i=1}^{n} K_h(x - x_i)(x - x_i), 
S_2 = \sum_{i=1}^{n} K_h(x - x_i)(x - x_i)(x - x_i)'
$$
\[
\hat{m}_0 = \sum_{i=1}^{n} K_h(x - x_i)Y_i, \quad \hat{m}_1 = \sum_{i=1}^{n} K_h(x - x_i)(x - x_i)Y_i. \\
\]

Then, by the usual partitioned inverse formula

\[
\hat{g}(x) = e_1' \left[ S_0 \quad S_1' \right]^{-1} \left( \hat{m}_0 \quad \hat{m}_1 \right) = (S_0 - S_1S_2^{-1}S_1)^{-1}(\hat{m}_0 - S_1S_2^{-1}\hat{m}_1)
\]

\[
= \frac{\sum_{i=1}^{n} a_iY_i}{\sum_{i=1}^{n} a_i}, \quad a_i = K_h(x - x_i)[1 - S_1S_2^{-1}(x - x_i)]
\]

It is straightforward though a little involved to find asymptotic approximations to the MSE. For simplicity we do this for scalar \( x \) case. Note that for \( g_0 = (g_0(x_1), \ldots, g_0(x_n))^t, \)

\[
\hat{g}(x) - g_0(x) = e_1'(X'WX)^{-1}X'W(Y - g_0) + e_1'(X'WX)^{-1}X'Wg_0 - g_0(x).
\]

Then for \( \Sigma = diag(\sigma^2(x_1), \ldots, \sigma^2(x_n)), \)

\[
E \left[ (\hat{g}(x) - g_0(x))^2 \right]_{x_1, \ldots, x_n} = e_1'(X'WX)^{-1}X'W\Sigma WX(X'WX)^{-1}e_1 + \left[ e_1'(X'WX)^{-1}X'Wg_0 - g_0(x) \right]^{2}
\]

An asymptotic approximation to MSE is obtained by taking the limit as \( n \) grows. Note that we have

\[
n^{-1}h^{-j}S_j = \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i)[(x - x_i)/h]^j
\]

Then, by the change of variables \( u = (x - x_i)/h, \)

\[
E \left[ n^{-1}h^{-j}S_j \right] = E \left[ K_h(x - x_i)((x - x_i)/h)^j \right] = \int K(u)u^j f_0(x - hu)du = \mu_j f_0(x) + o(1).
\]

for \( \mu_j = \int K(u)u^j du \) and \( h \to 0. \) Also,

\[
var(n^{-1}h^{-j}S_j) \leq n^{-1}E \left[ K_h(x - x_i)^2((x - x_i)/h)^{2j} \right] \leq n^{-1}h^{-1} \int K(u)^2u^{2j} f_0(x - hu)du \\
\leq Cn^{-1}h^{-1} \to 0
\]

for \( nh \to \infty. \) Therefore, for \( h \to 0 \) and \( nh \to \infty \)

\[
n^{-1}h^{-j}S_j = \mu_j f_0(x) + o_p(1).
\]

Now let \( H = diag(1, h). \) Then by \( \mu_0 = 1 \) and \( \mu_1 = 0 \) we have

\[
n^{-1}H^{-1}X'WXH^{-1} = n^{-1} \left[ \begin{array}{ccc} S_0 & h^{-1}S_1 \\ h^{-1}S_1 & h^{-2}S_2 \end{array} \right] = f_0(x) \left[ \begin{array}{ccc} 1 & 0 \\ 0 & \mu_2 \end{array} \right] + o_p(1).
\]
Next let $\nu_j = \int K(u)^2 u^j du$. Then by a similar argument we have

$$h \cdot \frac{1}{n} \sum_{i=1}^{n} K_h(x - x_i)^2[(x - x_i)/h] \sigma^2(x_i) = \nu_j f_0(x) \sigma^2(x) + o_p(1).$$

It follows by $\nu_1 = 0$ that

$$n^{-1}h^{-1}X'W \Sigma W XH^{-1} = f_0(x) \sigma^2(x) \begin{bmatrix} \nu_0 & 0 \\ 0 & \nu_2 \end{bmatrix} + o_p(1).$$

Then we have, for the variance term, by $H^{-1}e_1 = e_1$,

$$e_1'(X'WX)^{-1}X'W \Sigma W X(X'WX)^{-1}e_1 = n^{-1}h^{-1} \left( \frac{H^{-1}X'WXH^{-1}}{n} \right)^{-1} \frac{hH^{-1}X'W \Sigma W XH^{-1}}{n} \left( \frac{H^{-1}X'WXH^{-1}}{n} \right)^{-1} H^{-1}e_1$$

$$= n^{-1}h^{-1} \left( e_1' \left[ \begin{array}{cc} 1 & 0 \\ 0 & \mu_2 \end{array} \right]^{-1} \nu_0 \nu_1 \nu_2 \right) \left[ \begin{array}{cc} 1 & 0 \\ 0 & \mu_2 \end{array} \right]^{-1} e_1 \sigma^2(x) f(x) + o_p(1).$$

Assuming that $\mu_1 = 0$ as usual for a symmetric kernel we obtain

$$e_1'(X'WX)^{-1}X'W \Sigma W X(X'WX)^{-1}e_1 = n^{-1}h^{-1} \left( \nu_0 \sigma^2(x) f(x) + o_p(1) \right).$$

For the bias consider an expansion

$$g(x_i) = g_0(x) + g_0'(x)(x_i - x) + \frac{1}{2}g_0''(x)(x_i - x)^2 + \frac{1}{6}g_0'''(\bar{x}_i)(x_i - x)^3.$$

Let $r_i = g_0(x_i) - g_0(x) - [d g_0(x) / dx](x_i - x)$. Then by the form of $X$ we have

$$g = (g_0(x_1), \ldots, g_0(x_n))' = g_0(x)W e_1 - g_0'(x)W e_2 + r$$

It follows by $e_1'e_2 = 0$ that the bias term is

$$e_1'(X'WX)^{-1}X'W g - g_0(x) = e_1'(X'WX)^{-1}X'WX e_1 g_0(x) - g_0(x)$$

$$+ e_1'(X'WX)^{-1}X'WX e_2 g_0'(x) + e_1'(X'WX)^{-1}X'W r = e_1'(X'WX)^{-1}X'W r.$$

Recall that

$$n^{-1}h^{-j}S_j = \mu_j f_0(x) + o_p(1).$$

Therefore
\[
\begin{align*}
n^{-1}h^{-2}H^{-1}X'W((x - X_1)^2, \ldots, (x - X_n)^2)\frac{1}{2} &= \left( \begin{array}{cc}
n^{-1} & h^{-2} \\
n^{-1} & h^{-3} \end{array} \right) \begin{pmatrix} S_2 \\ S_3 \end{pmatrix} \frac{1}{2} g''_0(x) = f_0(x) \begin{pmatrix} \mu_2 \\ \mu_3 \end{pmatrix} \frac{1}{2} g''_0(x) + o_p(1).
\end{align*}
\]

Also, by \( g''_0(\bar{x}_i) \) bounded

\[
\left\| n^{-1}h^{-2}H^{-1}X'W \left( (x - x_1)^3 g''_0(\bar{x}_1), \ldots, (x - x_n)^3 g''_0(\bar{x}_n) \right)' \right\| \leq C \text{ max} \left\{ n^{-1}h^{-2} \sum_i K_h(x - x_i)|x - x_i|^3, n^{-1}h^{-2}S_4 \right\} \longrightarrow 0.
\]

Therefore, we have

\[
e'_1(X'WX)^{-1}X'Wr = h^2 e'_1 H^{-1} (H^{-1}X'WXH^{-1})^{-1} \frac{n}{n} h^{-2}H^{-1}X'Wr = \frac{h^2}{2} g''_0(x) e'_1 \begin{pmatrix} 1 & 0 \\ 0 & \mu_2 \end{pmatrix}^{-1} \begin{pmatrix} \mu_2 \\ \mu_3 \end{pmatrix} = \frac{h^2}{2} g''_0(x) \mu_2.
\]

Exercise: Apply analogous calculation to show kernel regression bias is

\[
\mu_2h^2 \left( \frac{1}{2} g''_0(x) + g''_0(x) \frac{f'_0(x)}{f_0(x)} \right)
\]

Notice bias is zero if function is linear.

Combining the bias and variance expression, we have the following form for asymptotic MSE:

\[
\frac{1}{nh} \nu_0 \sigma^2(x) f_0(x) + \frac{h^4}{4} g''_0(x)^2 \mu_2^2.
\]

In contrast, the kernel MSE is

\[
\frac{1}{nh} \nu_0 \sigma^2(x) f_0(x) + \frac{h^4}{4} \left[ g''_0(x) + 2g'_0(x) \frac{f'_0(x)}{f_0(x)} \right]^2 \mu_2^2.
\]

Bias will be much bigger near boundary of the support where \( f'_0(x)/f_0(x) \) is large. For example, if \( f_0(x) \) is approximately \( x^a \) for \( x > 0 \) near zero, then \( f'_0(x)/f_0(x) \) grows like \( 1/x \) as \( x \) gets close to zero. Thus, locally linear has smaller boundary bias. Also, locally linear has no bias if \( g_0(x) \) is linear but kernel obviously does.

Simple method is to take expected value of MSE.