Regression model

□ Consider a regression model with additive noise

\[ Y = f(X) + \varepsilon, \]

where \( \mathbb{E}(\varepsilon | X = x) = 0. \)

□ We have independent observations \((X_1, Y_1), \ldots, (X_n, Y_n)\) from that model.

Local average

□ Note that

\[ f(x) = \mathbb{E}(Y | X = x) \]

□ A local average (a.k.a. moving average) attempts to approximate this conditional expectation directly. It takes the form:

\[ \hat{f}(x) = \text{Ave}(Y_i | X_i \in N(x)) \]

where \( N(x) \) is a neighborhood of \( x \).

□ Note that there are two approximations here:

1. The expectation is approximated by an average.
2. The conditioning on \( X = x \) is approximated by conditioning on \( X \in N(x) \), where \( N(x) \) is a region around \( x \).
Choice of neighborhood type
The two main choices are:

- **$h$-ball neighborhood** where

  \[ N(x) = N_h(x) = \{ x' : |x' - x| \leq h \} \]

  This choice implies a constant window width, and this keeps the bias stable. Indeed, the bias comes from averaging over $N(x)$, a region around $x$ instead of averaging responses precisely at $x$.

- **$k$-nearest neighbors** where

  \[ N(x) = N_k(x) = \{ k \text{ closest points } X_i \text{'s to } x \} \]

  This choice implies a constant variance — assuming the errors have the same variance independent of the predictors.

Kernel regression (a.k.a. weighted local average)

- Choose a kernel function, often of the form

  \[ K_h(x, x_0) = D(|x - x_0|/h) \]

  where $D : \mathbb{R}_+ \rightarrow \mathbb{R}$ is non-increasing.

- The Nadaraya-Watson estimator based on that kernel is:

  \[ \hat{f}(x) = \frac{\sum_i K_h(x, X_i)Y_i}{\sum_i K_h(x, X_i)} \]

  The nearest neighbor version of this kernel estimator would be of the form:

  \[ \hat{f}(x) = \frac{\sum_i \mathbb{1}\{X_i \in N_k(x)\}K_h(x, X_i)Y_i}{\sum_i \mathbb{1}\{X_i \in N_k(x)\}K_h(x, X_i)} \]
Examples of Kernels

Our most basic requirement of $D$ is that it be non-increasing on $\mathbb{R}_+$. 

\begin{itemize}
  \item Uniform: $D(t) = \mathbb{I}\{t < 1\}$ [this leads to the local average]
  \item Triangle: $D(t) = (1 - t)_+$
  \item Epanechnikov: $D(t) = (1 - t^2)_+$
  \item Quartic: $D(t) = (1 - t^2)^2_+$
  \item TriCube: $D(t) = (1 - t^3)^3_+$ [used by the R function \texttt{loess}]
  \item Cosine: $D(t) = \cos\left(\frac{\pi}{2}t\right)\mathbb{I}\{t < 1\}$
  \item Gaussian: $D(t) = e^{-t^2/2}$ [also called heat kernel]
\end{itemize}

They are all supported on $[0, 1]$ except for the Gaussian kernel which is supported on the entire $\mathbb{R}^+$. However, the Gaussian kernel is fast-decaying.

Kernel methods are linear

\begin{itemize}
  \item Let $\hat{Y}_i = \hat{f}_h(X_i)$ be the usual fitted value for observation $i$. We have 
  \[
  \hat{Y}_i = \frac{\sum_r K_h(X_i, X_r)Y_r}{\sum_r K_h(X_i, X_r)} = \sum_r s_h(i, r)Y_r
  \]
  where 
  \[
  s_h(i, r) = \frac{K_h(X_i, X_r)}{\sum_t K_h(X_i, X_t)}
  \]
  Hence, 
  \[
  \hat{Y} = S_h Y
  \]
  where $Y = (Y_1, \ldots, Y_n)$ and $\hat{Y} = (\hat{Y}_1, \ldots, \hat{Y}_n)$, and 
  \[
  S_h = (s_h(i, r) : i, r \in \{1, \ldots, n\})
  \]
  is the smoother matrix.
  
  \item The degrees of freedom are defined as 
  \[
  df(h) = \text{trace}(S_h)
  \]
  This is in analogy with least squares, where $S$ is the hat matrix.
\end{itemize}
Local Linear Regression (LOESS)

- The local linear estimator is
  \[ \hat{f}_h(x) = \hat{\beta}_{h,0}(x) + \hat{\beta}_{h,1}(x)x \]
  where
  \[ (\hat{\beta}_{h,0}(x), \hat{\beta}_{h,1}(x)) = \arg \min_{b_0, b_1} \sum_{i=1}^{n} K_h(x, X_i) [Y_i - b_0 - b_1 X_i]^2 \]

- The estimate is linear in the response (with a different smoother matrix) and the degrees of freedom can be defined as before.

Local Polynomial Regression

- The local degree \( p \) polynomial estimator is
  \[ \hat{f}_h(x) = \hat{\beta}_{h,0}(x) + \hat{\beta}_{h,1}(x)x + \cdots + \hat{\beta}_{h,p}(x)x^p \]
  where
  \[ (\hat{\beta}_{h,0}(x), \ldots, \hat{\beta}_{h,p}(x)) = \arg \min_{b_0, \ldots, b_p} \sum_{i=1}^{n} K_h(x, X_i) [Y_i - b_0 - b_1 X_i - \cdots - b_p X_i^p]^2 \]

- The estimate is linear in the response (with a different smoother matrix) and the degrees of freedom can be defined as before.

Local Regression

- Suppose we assume a linear model in some basis \( \{g_0, \ldots, g_p\} \):
  \[ f_\theta(x) = \sum_{j=0}^{p} \theta_j g_j(x) \]

- The local linear estimator is \( \hat{f}_{h}(x) \), where
  \[ \hat{\theta}_h(x) = \arg \min_{\theta_0, \ldots, \theta_p} \sum_{i=1}^{n} K_h(x, X_i) [Y_i - f_\theta(X_i)]^2 \]

- The estimate is linear in the response (with a different smoother matrix) and the degrees of freedom can be defined as before.
Choosing of the tuning parameter

□ Assuming a model (when there is one) has been chosen. Then the window width $h$ (also called bandwidth) is the only tuning parameter.
   (This is replaced by the neighborhood size $k$ in the $k$-NN variant.)

□ This tuning parameter controls the degrees of freedom. The smaller $h$ is, the larger the degrees of freedom. The range is from 1 ($h \to \infty$) to $n$ ($h \to 0$) if all the $X_i$’s are distinct.

□ This parameter can be chosen to minimize an estimate of prediction error, for example, obtained by cross-validation.