Kernel regression

- Nonparametric regression. Suppose that (X_1, Y_1), …, (X_n, Y_n) are IID data and
$$Y_i = m(X_i) + \varepsilon_i \tag{1}$$
for $i = 1, \ldots, n$, where $(\varepsilon_1, \ldots, \varepsilon_n)$ is independent of $(X_1, \ldots, X_n)$, $E(\varepsilon_1) = 0$ and $Var(\varepsilon_1) = \sigma^2$. The problem of interest is to estimate m based on (X_1, Y_1), …, (X_n, Y_n).

- Kernel function. A kernel function k on $(-\infty, \infty)$ usually satisfies the usual constraints:
  (a) $k \geq 0$.
  (b) $\int_{-\infty}^{\infty} k(s)ds = 1$.
  (c) $\int_{-\infty}^{\infty} sk(s)ds = 0$.
  (d) $\int_{-\infty}^{\infty} s^2 k(s)ds < \infty$.

- Kernel regression estimator. Suppose that $(X_1, \ldots, X_n)$ is a random sample and X_i takes values in $(-\infty, \infty)$ for $i = 1, \ldots, n$. The kernel regression estimator for $m(x)$ with kernel k and bandwidth h is
$$\hat{m}(x) = \frac{\sum_{i=1}^{n} Y_i k\left(\frac{x - X_i}{h}\right)}{\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right)}.$$

- The estimation error $\hat{m}(x) - m(x)$.
$$\hat{m}(x) - m(x) = \frac{\underbrace{\frac{1}{nh}\sum_{i=1}^{n}(Y_i - m(x))k\left(\frac{x - X_i}{h}\right)}_{I}}{\underbrace{\frac{1}{nh}\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right)}_{II}}.$$

- Mean and variance of I.
$$\begin{aligned} E(I) &= E\left((nh)^{-1}\sum_{i=1}^{n}(Y_i - m(x))k((x - X_i)/h)\right) \\ &= h^2 \int u^2 k(u)du \left(\frac{f(x)m''(x)}{2} + f'(x)m'(x)\right) + o(h^2), \end{aligned}$$
and
$$\begin{aligned} Var(I) &= \frac{1}{nh^2}\left[E\left((Y_1 - m(x))^2 k^2((x - X_1)/h)\right)\right] \\ &\quad - \frac{h^2}{nh^2}\left[h^2 \int u^2 k(u)du \left(\frac{f(x)m''(x)}{2} + f'(x)m'(x)\right) + o(h^2)\right]^2 \end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{nh}\left[\sigma^2 f(x)\int k^2(u)du\right] - \frac{h^4}{n}\left[\int u^2k(u)du\left(\frac{f(x)m''(x)}{2} + f'(x)m'(x)\right)\right]^2 \\
&\quad + o\left(\frac{1}{nh}\right) + o\left(\frac{h^4}{n}\right).
\end{aligned}
$$

Thus if $nh \to \infty$ and $h \to 0$ as $n \to \infty$,

$$
\begin{aligned}
E(I^2) &= E\left((nh)^{-1}\sum_{i=1}^{n}(Y_i - m(x))k((x - X_i)/h)\right)^2 \\
&= \frac{1}{nh}\left[\sigma^2 f(x)\int k^2(u)du\right] + h^4\left[\int u^2k(u)du\left(\frac{f(x)m''(x)}{2} + f'(x)m'(x)\right)\right]^2 \\
&\quad + o\left(\frac{1}{nh}\right) + o(h^4).
\end{aligned}
$$

- Mean and variance for II. Let f be the density of X_i. Suppose that $nh \to \infty$ and $h \to 0$ as $n \to \infty$, then $E(II) - f(x) \to 0$ and $Var(II) \to 0$ as $n \to \infty$, so $II \approx f(x)$ for large n and $E(\hat{m}(x) - m(x))^2$ can be approximated by

$$
\frac{1}{nh}\left(\frac{\sigma^2 \int k^2(u)du}{f(x)}\right) + h^4\left[\int u^2k(u)du\left(\frac{m''(x)}{2} + \frac{f'(x)m'(x)}{f(x)}\right)\right]^2.
$$

- Kernel function on R^d. A kernel function k on R^d usually satisfies the usual constraints:

  (a) $k \geq 0$.

  (b) $\int k(s)ds = 1$.

  (c) $\int s_i k(s_1, \ldots, s_d)d(s_1, \ldots, s_d) = 0$ for $i = 1, \ldots, d$.

  (d) $\int \|s\|^2 k(s)ds < \infty$.

- Kernel regression estimator on R^d. Suppose that X_i takes values in R^d for $i = 1, \ldots, n$. The kernel regression estimator for $m(x)$ with kernel k and bandwidth h is

$$
\hat{m}(x) = \frac{\displaystyle\sum_{i=1}^{n} Y_i k\left(\frac{x - X_i}{h}\right)}{\displaystyle\sum_{i=1}^{n} k\left(\frac{x - X_i}{h}\right)}.
$$

- Bandwidth selection. We use leave-one-out cross validation to choose h for a given kernel k. Let $\hat{m}_{-i,h}$ be the kernel estimator for m with bandwidth h based on $(X_1, Y_1), \ldots, (X_{i-1}, Y_{i-1}), (X_{i+1}, Y_{i+1}), \ldots, (X_n, Y_n)$. Let

$$
RSSCV(h) = \sum_{i=1}^{n}(Y_i - \hat{m}_{-i,h}(X_i))^2.
$$

  Leave-one-out cross validation: choose the bandwidth h so that $RSSCV(h)$ is minimized.

- Bandwidth selection rule(s) can be found in [1].
- An example of approximating the bias of an estimator via simulation. Suppose that X_1, ..., X_n is a random sample from $N(\mu, 1)$ and consider estimating μ using the sample mean $\bar{X}$. The bias of $\bar{X}$ when $n = 50$ and $\mu = 20$ can be approximated using simulation.

```
#generate 1000 samples of size 50 from N(20,1)
#and store the 1000 sample means in x
x <- rep(0,1000)
for (i in 1:1000){ x[i] <- mean(rnorm(50,mean=20, sd=1)) }
#compute the approximate bias (expected value for sample mean - 20 )
sum(x - 20)/1000
```

- Exercise 1. Write a function using R with the following input and output:

  Input: data (X_1, Y_1), ..., (X_n, Y_n), kernel function k, bandwith h, and evaluation point x_0.

  Output: $\hat{m}(x_0)$.

  You may assume x_0 is one dimensional.

- Exercise 2. Consider the model in (1) with $n = 1000$, $m(x) = \sin(20x)$, where X_i is uniformly distributed on $[-1, 1]$ and the errors are IID $N(0, (0.01)^2)$. Let

$$I(x_0) = \left((nh)^{-1} \sum_{i=1}^{n} (Y_i - m(x_0))k((x_0 - X_i)/h) \right),$$

  where k is the probability density function of $N(0, 1)$.

  (a) Compute $E(I(x_0))$ for $x_0 = 0.1$ and $h \in \{0.01, 0.005, 0.001, 0.0005\}$. Note that

$$E(I(x_0)) = \int_{-\infty}^{\infty} (m(x_0 - hu) - m(x_0))f(x_0 - hu)k(u)du$$

  and the R command for computing $\int_a^b g(x)dx$ is `integrate(g,a,b)`.

  (b) Approximate $E(I(x_0))$ for $x_0 = 0.1$ and $h \in \{0.01, 0.0005\}$ by IID data (X_1, Y_1), ..., (X_n, Y_n) 10^4 times according to (1) with the above setup.

- Exercise 3. Write a function using R with the following input and output:

  Input: data (X_1, Y_1), ..., (X_n, Y_n), kernel function k, a vector of bandwidths $(h_1, \ldots, h_\ell)$ and evaluation point x_0.

  Output: $\hat{m}(x_0)$, where the bandwidth is chosen among h_1, ..., h_ℓ using leave-one-out cross validation.

  You may assume x_0 is one dimensional.

References

[1] W. Wand and M. Jones, *Kernel Smoothing*, Chapman & Hall, 1995.