Evalution of a nonparametric function estimtor

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

  (a) $X_1, \ldots, X_n$ are IID, or

  (b) $X_1, \ldots, X_n$ are not random.

  The problem of interest is to estimate m based on (X_1, Y_1), ..., (X_n, Y_n) for Case (a). For Case (b), m can be estimated well at some point x_0 only if there are enough X_is that are close to x_0.

- Suppose that the range of X_1 is $[a, b]$. Let $\hat{m}$ is an estimator of m. Then the integrated squared error (ISE) is
$$\int_a^b (\hat{m}(x) - m(x))^2 dx$$
and one can use the integrated mean squared error (IMSE) to evaluation the performance of $\hat{m}$.
$$\text{IMSE } = \int_a^b E(\hat{m}(x) - m(x))^2 dx = E(\text{ISE}). \tag{2}$$
  - ISE can be computed using the R command `integrate`.
  - To approximate IMSE, one needs to generate IID N data sets data_1, ..., data_N from (1) and let E_j be the ISE for the $\hat{m}$ computed based on data_j. Then
$$\text{IMSE} = E(\text{ISE}) \approx \frac{1}{N} \sum_{j=1}^{N} E_j \tag{3}$$
for large N.

- The R command `integrate(g,a,b)` computes $\int_a^b g(x)dx$. Note that g must accept a vector input.

  Example 1. Find $\int_0^1 \left(\int_0^x \sin(y^2) dy \right) dx$.

```
f1 <- function(y){ sin(y^2) }
g <- function(x){ integrate(f1, 0, x)$value }
g1 <- Vectorize(g)
integrate(g1, 0, 1)$value
```

  - Note that running `integrate(g, 0, 1)` gives an error.

- Recall that
$$RSSCV(h) = \sum_{i=1}^{n} (Y_i - \hat{m}_{-i,h}(X_i))^2,$$
where
$$E(Y_i - \hat{m}_{-i,h}(X_i))^2 = E \int (m(x) - \hat{m}_{-i,h}(x))^2 f_X(x)dx + \sigma^2,$$
and f_X is the density of X_i. We expect
$$\begin{aligned} \frac{RSSCV(h)}{n} - \sigma^2 &\approx E \int (m(x) - \hat{m}_{-i,h}(x))^2 f_X(x)dx \\ &\approx E \int (m(x) - \hat{m}(x))^2 f_X(x)dx. \end{aligned}$$
When the distribution of X_i is $Uniform(0,1)$, we expect $RSSCV/n - \sigma^2$ to be close to IMSE.

- Exercise 1. Let $N = 100$. Simulated N data sets from (1) with $m(x) = \sin(20x)$, $n = 1000$, $X_1, \ldots, X_n$ are IID $Uniform(0,1)$, $\varepsilon_1, \ldots, \varepsilon_n$ are IID $N(0, \sigma^2)$ errors with $\sigma = 0.05$. Compute the IMSE using (3) for the kernel regression estimator with the bandwidth $h \in \{0.005, 0.01, 0.1\}$.

- Exercise 2. Let $k = 10$. Generate k data sets from the model in (1) with $m(x) = \sin(20x)$, $n = 1000$, $X_1, \ldots, X_n$ are IID $Uniform(0,1)$, $\varepsilon_1, \ldots, \varepsilon_n$ are IID $N(0, \sigma^2)$ errors with $\sigma = 0.05$.

  (a) Compute $RSSCV/n - \sigma^2$ for each data set for $h \in \{0.1, 0.01\}$. Can we conclude that all the k values for $RSSCV/n - \sigma^2$ are close to the IMSE values for $h \in \{0.1, 0.01\}$ from Exercise 1?

  (b) Suppose that the k data sets are generated the same way as in Part (a) except that the distribution for each X_i is the beta distribution $beta(2,2)$. Compute $RSSCV/n - \sigma^2$ for each data set for $h \in \{0.1, 0.01\}$. Can we conclude that all the k values for $RSSCV/n - \sigma^2$ are close to the IMSE values for $h \in \{0.1, 0.01\}$ from Exercise 1?

  (c) If we define IMSE* for $\hat{m}$: an estimator of m to be
  $$E \int (\hat{m}(u) - m(u))^2 f_X(u)du,$$
  where f_X is the density for the beta distribution $beta(2,2)$. Can we conclude that all the k values for $RSSCV/n - \sigma^2$ from Part (b) are close to the IMSE* values with $h \in \{0.1, 0.01\}$? You may approximate the IMSE* values using averages over 100 weighted ISE values.

  - Note. The R command `rbeta(n, 2,2)` generates `n` random numbers from $beta(2,2)$.

- Exercise 3. Let $N = 100$. Simulated N data sets from (1) with $m(x) = \sin(20x)$, $n = 1000$, $X_1, \ldots, X_n$ are IID $Uniform(0,1)$, $\varepsilon_1, \ldots, \varepsilon_n$ are IID $N(0, \sigma^2)$ errors with $\sigma = 0.05$. Compute the IMSE using (3) for the kernel regression estimator with the bandwidth chosen using leave-one-out cross validation, where the bandwidth h is in $\{0.005, 0.01, 0.1\}$.