EXERCISE 14

WEIYU LI

1. Suppose $\{X_i, Y_i\}$ are bivariate random samples with

$$Y_i = m(X_i) + u_i,$$

where $m(\cdot)$ is an unknown smooth function, and u_i satisfies $E[u_i|X_i] = 0, Var(u_i|X_i) = \sigma^2(X_i), a.s.$

(1) Solve the local linear estimation of $m(x)$, and its asymptotic bias and variance (main terms).

(2) Solve the local linear estimation $\hat{m}_{ll}^{(1)}(x)$ of the first derivative of $m(x)$, and prove that

$$\hat{m}_{ll}^{(1)}(x) = \frac{\sum_{i=1}^n (Y_i - \bar{Y}_k)(X_i - \bar{X}_k)K_{i,x}}{\sum_{i=1}^n (X_i - \bar{X}_k)^2 K_{i,x}},$$

$$\hat{m}_{ll}(x) = \bar{Y}_k - (\bar{X}_k - x)\hat{m}_{ll}^{(1)}(x),$$

where $\bar{Y}_k = \sum_{i=1}^n Y_i K_{i,x} / \sum_{i=1}^n K_{i,x}$, $\bar{X}_k = \sum_{i=1}^n X_i K_{i,x} / \sum_{i=1}^n K_{i,x}$ and $K_{i,x} = K_h(x - X_i)$.

Solve. We are going to solve those two problems together, since the notation $\hat{m}_{ll}(x)$ in (2) is exactly what we focus on in (1). With kernel K and bandwidth h, consider the minimization problem

$$\min_{m,\beta} \sum_{i=1}^n (Y_i - m - (X_i - x)\beta)^2 K_{i,x},$$

where $K_{i,x} = K_h(x - X_i) = \frac{1}{h}K(\frac{x - X_i}{h})$.

Let $M(x) = \begin{pmatrix} m(x) \\ \beta(x) \end{pmatrix}$, $X = \begin{pmatrix} 1 & X_1 - x \\ \vdots & \vdots \\ 1 & X_n - x \end{pmatrix}$, $W = diag(K_{1,x}, \cdots, K_{n,x})$ and $Y = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}$, then the minimization problem can be regarded as a least square problem with solution

$$\hat{M}(x) = (X'WX)^{-1}X'WY := \begin{pmatrix} \hat{m}_{ll}(x) \\ \hat{m}_{ll}^{(1)}(x) \end{pmatrix},$$

which consists of the local linear estimations of $m(x)$ and its first derivative.

We can explicitly write

$$\begin{aligned} X'WX &= \begin{pmatrix} \sum_i K_{i,x} & \sum_i (X_i - x)K_{i,x} \\ \sum_i (X_i - x)K_{i,x} & \sum_i (X_i - x)^2 K_{i,x} \end{pmatrix} \\ &= \sum_i K_{i,x} \begin{pmatrix} 1 & \bar{X}_k - x \\ \bar{X}_k - x & \sum_i (X_i - x)^2 K_{i,x} / \sum_i K_{i,x} \end{pmatrix}. \end{aligned}$$

Date: 2019/11/25.

liweiyu@mail.ustc.edu.cn.

Further noticing that

$$\begin{aligned}
&\sum_i (X_i - x)^2 K_{i,x} \\
=&\sum_i \left[(X_i - \bar{X}_k) + (\bar{X}_k - x)\right]^2 K_{i,x} \\
=&\sum_i (X_i - \bar{X}_k)^2 K_{i,x} + 2(\bar{X}_k - x)\sum_i (X_i - \bar{X}_k) K_{i,x} + (\bar{X}_k - x)^2 \sum_{i=1}^n K_{i,x} \\
=&\sum_i (X_i - \bar{X}_k)^2 K_{i,x} + (\bar{X}_k - x)^2 \sum_{i=1}^n K_{i,x}
\end{aligned}$$

simplifies the matrix as

$$X'WX = \sum_i K_{i,x} \begin{pmatrix} 1 & \bar{X}_k - x \\ \bar{X}_k - x & \frac{\sum_i (X_i - \bar{X}_k)^2 K_{i,x}}{\sum_i K_{i,x}} + (\bar{X}_k - x)^2 \end{pmatrix}.$$

Similarly, we have

$$\begin{aligned}
X'WY &= \begin{pmatrix} \sum_i Y_i K_{i,x} \\ \sum_i (X_i - x) Y_i K_{i,x} \end{pmatrix} \\
&= \sum_i K_{i,x} \begin{pmatrix} \bar{Y}_k \\ \sum_i (X_i - x) Y_i K_{i,x} / \sum_i K_{i,x} \end{pmatrix} \\
&= \sum_i K_{i,x} \begin{pmatrix} \bar{Y}_k \\ \frac{\sum_i (Y_i - \bar{Y}_k)(X_i - \bar{X}_k) K_{i,x}}{\sum_i K_{i,x}} + \bar{Y}_k (\bar{X}_k - x) \end{pmatrix}.
\end{aligned}$$

A trick we use is to eliminate part of the second rows that is $(\bar{X}_k - x)$-proportional to the first rows. In specific, multiplying the nonsingular $A = (\sum_i K_{i,x})^{-1} \begin{pmatrix} 1 & 0 \\ -(\bar{X}_k - x) & 1 \end{pmatrix}$, we have

$$\begin{aligned}
AX'WX &= \begin{pmatrix} 1 & \bar{X}_k - x \\ 0 & \frac{\sum_i (X_i - \bar{X}_k)^2 K_{i,x}}{\sum_i K_{i,x}} \end{pmatrix}, \\
AX'WY &= \begin{pmatrix} \bar{Y}_k \\ \frac{\sum_i (Y_i - \bar{Y}_k)(X_i - \bar{X}_k) K_{i,x}}{\sum_i K_{i,x}} \end{pmatrix}.
\end{aligned}$$

Consequently,

$$\begin{aligned}
\begin{pmatrix} \hat{m}_{ll}(x) \\ \hat{m}_{ll}^{(1)}(x) \end{pmatrix} &= \left(AX'WX\right)^{-1} AX'WY \\
&= \begin{pmatrix} 1 & \bar{X}_k - x \\ 0 & \frac{\sum_i (X_i - \bar{X}_k)^2 K_{i,x}}{\sum_i K_{i,x}} \end{pmatrix}^{-1} \begin{pmatrix} \bar{Y}_k \\ \frac{\sum_i (Y_i - \bar{Y}_k)(X_i - \bar{X}_k) K_{i,x}}{\sum_i K_{i,x}} \end{pmatrix} \\
&= \begin{pmatrix} 1 & -(\bar{X}_k - x) \frac{\sum_i K_{i,x}}{\sum_i (X_i - \bar{X}_k)^2 K_{i,x}} \\ 0 & \frac{\sum_i K_{i,x}}{\sum_i (X_i - \bar{X}_k)^2 K_{i,x}} \end{pmatrix} \begin{pmatrix} \bar{Y}_k \\ \frac{\sum_i (Y_i - \bar{Y}_k)(X_i - \bar{X}_k) K_{i,x}}{\sum_i K_{i,x}} \end{pmatrix}
\end{aligned}$$

$$= \begin{pmatrix} \bar{Y}_k - (\bar{X}_k - x)\frac{\sum_i (Y_i-\bar{Y}_k)(X_i-\bar{X}_k)K_{i,x}}{\sum_i (X_i-\bar{X}_k)^2 K_{i,x}} \\ \frac{\sum_i (Y_i-\bar{Y}_k)(X_i-\bar{X}_k)K_{i,x}}{\sum_i (X_i-\bar{X}_k)^2 K_{i,x}} \end{pmatrix},$$

which completes the proof in (2).

From the theorem in Page 27 of *Lec14.pdf*, asymptotic bias and variance of $\hat{m}_{ll}(x)$ are

$$bias\left(\hat{m}_{ll}(x)\right) = \frac{\kappa_{21}}{2}h^2 m''(x),$$
$$Var\left(\hat{m}_{ll}(x)\right) = \frac{\kappa_{02}\sigma^2(x)}{nhf(x)},$$

where $f(\cdot)$ is the density of X_1. □

2. Consider the exponential generalized linear model

$$Y|X = x \sim Exp(\lambda(x)), \lambda(x) = e^{\beta_0+\beta_1 x}.$$

Using local likelihood estimation, write an estimate function depending on the sample X, Y, estimate point x, bandwidth h and kernel K. Generate a simulated dataset, use your function to estimate, and choose the optimal bandwidth by cross-validation.

Solve. Note that in local regression, the model has the expression

$$Y_i = m(X_i) + \epsilon_i,$$

where $m(x) = E(Y|X = x) = \frac{1}{\lambda(x)} = e^{-\beta_0-\beta_1 x}$. On the other hand, the link function is $g(\mu) = -\log(\mu)$ in GLM fitting. The log-likelihood of β is

$$l(\beta) = \sum_{i=1}^{n} \log\left(f(y)\right) = \sum_{i=1}^{n} \log\left(\lambda(X_i)\right) - \lambda(X_i)Y_i.$$

The local log-likelihood of β around x is then

$$\begin{aligned} l_{x,h}(\beta) =& \sum_{i=1}^{n} \left[\log(\lambda(X_i - x)) - \lambda(X_i - x)Y_i\right] K_h(x - X_i) \\ =& \sum_{i=1}^{n} \left[\beta_0 + \beta_1(X_i - x) - e^{\beta_0+\beta_1(X_i-x)}Y_i\right] K_h(x - X_i). \end{aligned}$$

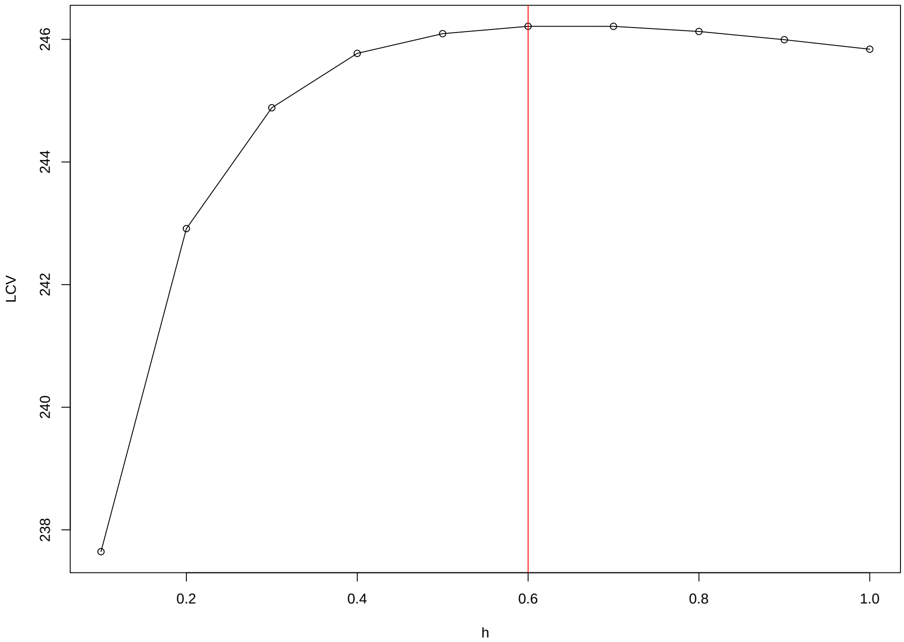
Maximizing $l_{x,h}(\beta)$ yields $\hat{\beta} = (\hat{\beta}_0, \hat{\beta}_1)'$, thus we have $\hat{m}_l(x; h, p) = g^{-1}(\hat{\beta}_0) = e^{-\hat{\beta}_0}$. For optimal bandwidth, use cross-validation, that is, maximize

$$LCV(h) = \sum_{i=1}^{n} l(Y_i, \hat{\lambda}_{-i}(X_i)).$$

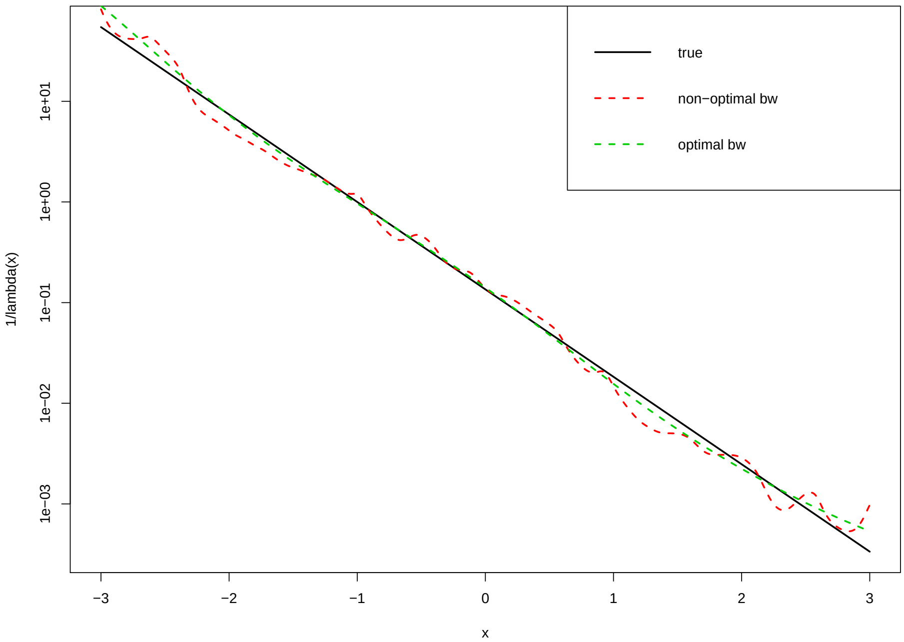
From the above theoretical analysis, we provide the codes following the material *Lec14.r*:

```
## Initialization
n <- 200
truebeta <- c(2,2) #true beta_0 and beta_1
lambda <- function(x) exp(truebeta[1] + truebeta[2] * x)
lik <- function(y,beta) beta - exp(beta) * y
# likelihood function with $\lambda=e^\hat{\beta}$
set.seed(0)
```

```
## Generate a dataset
X <- runif(n = n, -3, 3)
Y <- rexp(n = n, rate = lambda(X))

## Set bandwidth and evaluation grid
h <- 0.1
x <- seq(-3, 3, l = 501)

## Optimize the weighted log-likelihood explicitly
suppressWarnings(
  fitNlm <- sapply(x, function(x) {
    K <- dnorm(x = x, mean = X, sd = h)
    nlm(f = function(beta) {
      sum(K * (Y * exp(beta[1] + beta[2] * (X - x))
               - (beta[1] + beta[2] * (X - x))))
    }, p = c(0, 0))$estimate[1]
  })
)

## Exact LCV
h <- seq(0.1,1, by = .1)
suppressWarnings(
  LCV <- sapply(h, function(h) {
    sum(sapply(1:n, function(i) {
      K <- dnorm(x = X[i], mean = X[-i], sd = h)
      lik(Y[i],nlm(f = function(beta) {
        sum(K * (Y[-i] * exp(beta[1] + beta[2] * (X[-i] - X[i]))
                 - ( beta[1] + beta[2] * (X[-i] - X[i]))))
      }, p = c(0, 0))$estimate[1])
    }))
  })
)
plot(h, LCV, type = "o")
abline(v = h[which.max(LCV)], col = 2)

## Compare the optimal bandwidth with the non-optimal one
h <- h[which.max(LCV)]
suppressWarnings(
  fitNlm.opt <- sapply(x, function(x) {
    K <- dnorm(x = x, mean = X, sd = h)
    nlm(f = function(beta) {
      sum(K * (Y * exp(beta[1] + beta[2] * (X - x))
               - (beta[1] + beta[2] * (X - x))))
    }, p = c(0, 0))$estimate[1]
  })
)
```




```
plot(x, 1/lambda(x),  type = "l", lwd = 2, log = "y")
lines(x,exp(-fitNlm), col = 2, lwd = 2, lty = 2) #inverse of link function g
lines(x,exp(-fitNlm.opt), col = 3, lwd = 2, lty = 2)
legend("topright",legend = c("true", "non-optimal bw", "optimal bw"),
       col = 1:3, lwd = c(2, 2, 2), lty = c(1, 2, 2))
```



We can observe that optimal bandwidth gives better fitting. □