EXERCISE 11

WEIYU LI

1. For the dataset *faithful* in R, test whether *eruptions* and*waiting* are independent.

```
### initialize
library(ks)
set.seed(0)
alpha <- 0.05 # alpha for the test
x.true <- as.numeric(faithful[,1]) # x is for eruptions
y.true <- as.numeric(faithful[,2]) # y is for waiting
n <- length(x.true) # number of samples
B <- 20 # bootstrap time

# function to calculate T with sample (x,y)
Tstat <- function(x, y){ # return T
  n <- length(x) # number of samples
  hx <- hpi(x) # bandwidth for x
  hy <- hpi(y) # bandwidth for y
  fx <- rep(0, n)
  fy <- rep(0, n)
  # leave-one-out kernel estimators of f(x),f(y)
  Kx <- matrix(rep(0, n^2),n)
  Ky <- matrix(rep(0, n^2),n)
  # The (i,j)-th element of Kx is K_hx(Xj-Xi)
  # Compute K: Gaussian kernel
  index <- 1:n
  for (i in index) {
    for (j in index[-i]) {
      Kx[i,j] <- dnorm((x[j] - x[i]) / hx)/ hx
      Ky[i,j] <- dnorm((y[j] - y[i]) / hy)/ hy
    }
  }
  fx <- n / (n-1) * colMeans(Kx)
  fy <- n / (n-1) * colMeans(Ky)
  I <- n / (n-1) * mean(Kx * Ky) + mean(fx %*% t(fy)) - 2 * mean(fx * fy)
  sigma <- sqrt(2 * sum(Kx^2 * Ky^2) / (n^2 * hx * hy))
  T <- n * sqrt(hx * hy) * I / sigma
  return(T)
}
```

Date: 2019/11/11.

liweiyu@mail.ustc.edu.cn.

```
### (1) leave-one-out method
T.true <- Tstat(x.true, y.true)
if (abs(T.true) > qnorm(1 - alpha / 2)){
  cat('Using leave-one-out, we reject independent structure.')
} else{
  cat('Using leave-one-out, we accept independent structure.')
}
# result: Using leave-one-out, we reject independent structure.

### (2) Boostrap
Tstar <- rep(0,B)
for (i in 1:B) {
  xstar <- sample(x.true, n, replace = TRUE)
  ystar <- sample(x.true, n, replace = TRUE)
  Tstar[i] <- Tstat(xstar, ystar)
} # Bootstrap
p.boot <- mean(T.true < Tstar)
if (p.boot < alpha){
  cat('Using bootstrap, we reject independent structure.')
} else{
  cat('Using bootstrap, we accept independent structure.')
}
# result: Using bootstrap, we reject independent structure.

### (3) Compare the two methods
cat('p-value for true T statistics is: ', pnorm(-abs(T.true)) * 2)
# 6.373953e-32
cat('p-value for the bootstrap method is: ', p.boot)
# 0
```

We conduct two methods: (i) computing T statistics based on the true data, and construct a normal test; (ii) bootstrap the data, and compute the bootstrap p-value. Both the two methods give extremely small p-values, so we confidently **reject** the independent hypothesis (null hypothesis.)

Remark 1. *Intuitively, the dataset has two clusters, so the two variables are highly correlated. Therefore, we are subject to believe they are dependent, which results in setting independent structure as the null hypothesis.*

□