Advanced Statistical Modeling

Non-parametric models - Alternative estimations of conditional variance

Haoran Mo, Alexandra Yamaui

November 22, 2017

In this exercise we are going estimate the residual variance of the percentage of the lower status of the Boston population (LSTAT) of the boston R dataset respect to average number of rooms per dwellin (RM). We will use methods that not require a previous estimation of the regression function.

The first approach is Rice. We would like to make it more clear: The variance of a random variable tells us something about the spread of the possible values of the variable. For a discrete random variable $y_i - y_{i-1}$, the variance of $y_i - y_{i-1}$ is written as $\text{Var}(y_i - y_{i-1})$, and $\text{Var}(y_i - y_{i-1}) = E[((y_i - y_{i-1}) - m)^2]$ where m is $E(y_i - y_{i-1})$, it can be written as $Var(y_i - y_{i-1}) = E[(y_i - y_{i-1})^2] - m^2$. After further more steps as shown in the chapter 2.4, we can obtain the function $\hat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} (y_i - y_{i-1})^2$

The second approach is the method proposed by Gasser, Stroka and Jennen-Steinmetz in 1986, which consists in linear interpolation of every point (x_i, y_i) with the previous and following observations (x_{i-1}, y_{i-1}) and (x_{i+1}, y_{i+1}), respectively. The observations are sorted based on the value of x_i in ascending order. The idea behind this approach is that $\hat{y_i}$ $(\hat{m}_i)$ is approximetly equal to $(x_{i-1}, m(x_{x-i}))$ and $(x_{x+1}, m(x_{x+i}))$ if the function m is smooth and x_i, x_{i-1} and x_{i+1} are close enough.

The linear interpolation for x_i is defined by:

$$\hat{y}_i = \frac{x_{i+1} - x_i}{x_{i+1} - x_{i-1}} y_i + \frac{x_i - x_{i-1}}{x_{i+1} - x_{i-1}} y_{i+1} = a_i y_{i-1} + b y_{i+1}$$

The estimation of the residuals would be the difference between the estimated value $\hat{y}_i$ from the interpolation and the real value y_i with an expected value $E(\tilde{\varepsilon}) \approx 0$:

$$\tilde{\varepsilon} = \hat{y}_i - y_i = a_i y_{i-1} + b y_{i+1} - y_i$$

The residuals can be seen as the deviation from the true value, therefore:

$$E(\tilde{\varepsilon}^2) \approx V(\tilde{\varepsilon}_i) = (a_i^2 + b_i^2 + 1)\sigma^2$$

and the residual variance $\hat{\sigma}^2$ can be approximated as:

$$\hat{\sigma}^2 = \frac{1}{n-2} \sum_{i=2}^{n-1} \frac{1}{a_i^2 + b_i^2 + 1} \tilde{\varepsilon}_i^2$$

Below we present the implementation of a function that calculates the estimation of the residual variance through both approaches. For the second method, in the cases when the previous (x_{i-1}) and following (x_{i+1}) observations are equal, the maximum of lower observations and the minimum of higher observations are taken instead.

```
calculate_residual_variance <- function(X,Y) {
  XY <- data.frame(X,Y)
  XY <- XY[order(XY$X),]
  rownames(XY) <- 1:nrow(XY)
  n <- length(Y)

  # Rice
```

```
  t2 = XY[-c(length(X)),"Y"]
  t1 = XY[-c(1),"Y"]
  rice.sigma_2 <- (sum((t2-t1)^2))/(2*(n-1))

  # Gasser, Sroka, and Jennen-Steinmetz
  summation <- 0

  for (i in 2:(n-1)) {
    xi <- XY[i, 'X']

    x.previous = XY[i-1,'X'] # x_i-1
    x.following = XY[i+1,'X'] # x_i+1
    y.previous = XY[i-1,'Y'] # y_i-1
    y.following = XY[i+1,'Y'] # y_i+1

    if (x.previous == x.following) {
      x.following <- min(XY[XY$X > xi,]$X)
      x.previous <- max(XY[XY$X < xi,]$X)
    }

    a_i <- (x.following - xi)/(x.following - x.previous)
    b_i <- (xi - x.previous)/(x.following - x.previous)

    y.hat_i <- a_i*y.previous + b_i*y.following

    residual.hat_i <- y.hat_i - XY[i, 'Y']
    summation <- summation + (residual.hat_i^2/(a_i^2 + b_i^2 + 1))
  }

  gasser.sigma2 <- summation/(n - 2)
  return(list(rice.sigma_2, gasser.sigma2))
}

X <- boston.c$LSTAT
Y <- boston.c$RM
(result <- calculate_residual_variance(X,Y))
```

Lastly, we are going to compare the estimated values using other R packages methods. The first one is called Local Polynomial Regression Fitting (loess) and it fits a polynomial surface determined by one or more numerical predictors, using local fitting, and a second method described in the book of Bowman & Azzalini (1997), that implements a nonparametric smoothing method. In the Table1 we do a comparisson of the estimated values from the four methos, which returned similar results.

```
loess.fit <- loess(RM~LSTAT, data = boston.c)
residual.variance.loess <- var(loess.fit$residuals)
sigma.loess <- sqrt(residual.variance.loess)
sm.fit <- sm.regression(boston.c$LSTAT,boston.c$RM)
sigma.sm <- sm.fit$sigma
```

Table 1: Residual variance estimation

Method	Residual.variance
Rice	0.5315709
Gasser	0.5174054

Method	Residual.variance
loess	0.5008800
sm.regression	0.5097599

We can see that the four estimations are very similar, which indicates the correctness of the Rice and Gasser methods implemented by ourselves.