Report on exercise 4.2: Conditional Residual Variance Estimation

Martin Guy, Hannes Leskelä

October 24, 2016

Abstract

In this exercise, we want to compare alternative residual variance estimators. These ones do not require a previous estimation of the regression function. The working material that will be used is **Boston housing Data**. In order to check if these estimators are "good", we will compare them with estimators from `loess` and `sm.regression`.

Differents proposals

Proposal of Rice

In a paper from 1984, Rice suggests that if we consider a non-parametric regression model:

$$y_i - y_{i-1} = m(x_i) - m(x_{i-1}) + (\epsilon_i - \epsilon_{i-1}), i = 1, ..., n$$

and assumed that the function m is smooth and that x_i and x_{i-1} are sufficiently close, the variance $\widehat{\sigma}^2$ could be expressed as

$$\widehat{\sigma}^2 = \frac{1}{2(n-1)} \sum_{i=2}^{n} (y_i - y_{i-1})^2$$

Proposal of Gasser, Sroka and Jennen-Steinmetz (1986)

The estimation of σ for this proposal is:

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

where

$$a_i = \frac{x_{i+1} - x_i}{x_{i+1} - x_{i-1}}$$

$$b_i = \frac{x_i - x_{i-1}}{x_{i+1} - x_{i-1}}$$

$$\tilde{\epsilon}_i = a_i y_{i-1} + b_i y_{i+1} - y_i$$

It can happen in real datasets that we find 3 (or more) consecutive repeated values for the explanatory variable x. In this case, $x_{i+1} = x_{i-1}$ and then a_i and b_i are not well defined. In order to avoid that, we define $x_{u(i)}$ as the lowest value of x_j among those being greater than x_i and $x_{l(i)}$ as the largest value of x_j among those being lower than x_i. If such a case happens, we replace the denominator $x_{i+1} - x_{i-1}$ by $x_{u(i)} - x_{l(i)}$.

Comparison of estimators

Estimation method	**Estimation of σ**
`loess`	0.503
`sm.regression`	0.510
Rice	0.498
Gasser et al.	0.458

We can observe that the estimates are quite near the `loess` or `sm.regression` models. However, Rice and Gasser et al. are "nice" methods since they do not require a previous estimation of the regression function to calculate the variance. This means that we can estimate the variance of a model without first making the regression function, and thus decide if it is worth doing a regression model or not depending on this value.