Indian Institute of Technology, Kanpur

Data Science Lab
(MTH312A)
HOME WORK-2
REPORT

Group Members:
Mrinmoy Saha (221352)
Sudesh Kumari (221440)
Vijay Soren (211163)
Manjeet Chaudhary (221346)
Pravin Raman Bharti (221375)

Question 1: Download a two sample multivariate data, where dimension of the data is larger than sample size of the data. Check whether the distributions associated with two samples are independent or not.

Answer:

About Data Set: The breast cancer dataset contains features computed from images of breast mass fine needle aspirates. It includes features like radius, texture, perimeter, area, smoothness, compactness, concavity, concave points, symmetry, and fractal dimension. Based on these features, the data categorizes whether a tumor is benign (B) or malignant (M).

Theory: We check the independence of the two data sets which are based on Malignant (M) and Benign (B) using the approach of Empirical cumulative distribution function (ECDF).

The empirical distribution function $F_n(t)$ is defined as:

$$Fn(t) = \frac{\text{number of elements in the sample} \leq t}{n} = \frac{1}{n}\sum i = 1^n \mathbf{1}_{X_i \leq t}$$

Mathematically we are checking whether

$$F_n(\underline{X}, \underline{Y}) = F_n(\underline{X}) \cdot F_n(\underline{Y})$$

Interpretation:

We define functions to calculate ECDFs for the joint data and the data based on benign (B) and malignant (M) tumors. Then based on a range of values we check whether ECDF for the joint data is the product of the ECDFS for Benign (B) and Malignant (M) or not. So we calculated the absolute distance between the ECDF for the joint data and the product of the ECDFS for Benign (B) and Malignant (M). Now we take the supremum of the absolute distance and determine the independence of the two tumors based on this value.

Conclusion:

We see that the supremum distance came out to be 0.63. Since this value is significantly greater than 0, we conclude that the data based on the two categories Benign (B) and Malignant (M) are dependent.

Question 2: Download a data, which is suitable for non-parametric regression models. For this data, estimate the regression function and its first and second derivatives using local polynomial mean and median approach. Compare the performance of the estimators obtained from both approaches.

Answer:

Local polynomial regression is a non-parametric regression technique used to model the relationship between a dependent variable and one or more independent variables. It's particularly useful when the relationship between the variables is not linear and may change over different regions of the independent variable(s).

In local polynomial regression, instead of fitting a single global regression model to the entire dataset, separate regression models are fit to different subsets of the data. These subsets, or "neighborhoods," are defined based on the values of the independent variable(s) around a particular point of interest. The size of the neighborhood is determined by a bandwidth parameter.

About the Data Set:

We use the Boston data set from the MASS library of R and we take two features: LSTAT, which indicates a percentage of lower status of population, and MEDV, which indicates the median value of owner-occupied homes in $1000's$.

Theory:

Here Taylor's series is used for the construction of the Local Polynomial Regression. The Taylor series expansion of a function $f(x)$ about a point a is given by:

$$\begin{aligned} f(x) = f(a) + f'(a)(x-a) + \frac{f''(a)}{2!}(x-a)^2 \\ + \frac{f'''(a)}{3!}(x-a)^3 + \cdots \\ + \frac{f^{(n)}(a)}{n!}(x-a)^n \end{aligned}$$

In local polynomial regression, a low-order weighted least squares (WLS) regression is fit at each point of interest, x, using data from some neighborhood around x. Let (X_i, Y_i) be pairs of data points such that

$$Y_i = m(X_i) + \sigma(X_i)\varepsilon_i$$

where $\varepsilon_i \sim \mathcal{N}(0,1)$, $\sigma^2(X_i)$ is the variance of Y_i at the point X_i, and X_i comes from some distribution, f. In some cases, homoskedastic variance is assumed, so we let $\sigma^2(X) = \sigma^2$. It is typically of interest to estimate $m(x)$.

$$m(x) \approx m(x_0) + m^{(1)}(x_0)(x-x_0) + \cdots + \frac{m^{(p)}(x_0)}{p!}(x-x_0)^p$$

We can estimate these terms using weighted least squares. Minimize:

$$\sum_{i=1}^{n} \left(Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \right)^2 K_h(X_i - x_0).$$

h controls the size of the neighborhood around x_0, and $K_h(\cdot)$ controls the weights, where

$$K_h(\cdot) \equiv \frac{1}{h} K\left(\frac{\cdot}{h}\right),$$

and K is a kernel function and h is bandwidth.

For the median case, we go with the same approach, but here we try to minimize:

$$\sum_{i=1}^{n} \left| Y_i - \sum_{j=0}^{p} \beta_j (X_i - x_0)^j \right| K_h(X_i - x_0).$$

Interpretation:

Here we see that there is no linear relationship between the two variables, so the data is suitable for non-parametric regression from the figure 1 in the next page.

The R-squared (R^2) values and mean squared error (MSE) for the mean and median approaches in local polynomial regression are calculated.

Mean squared error (MSE) for the mean and median approaches in local polynomial regression are calculated.

- For the mean approach:

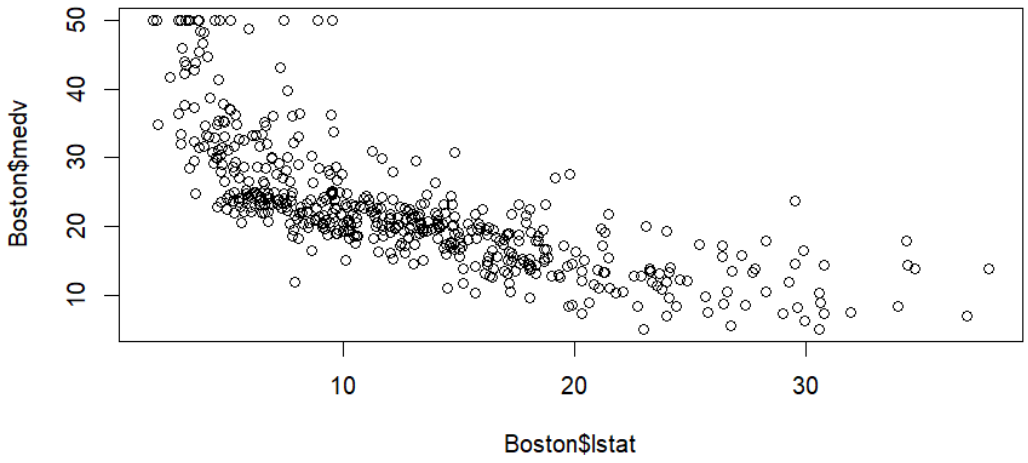

Figure 1: Scatter Plot of our Data

  - MSE: 12846.58

- For the median approach:

  - MSE: 13896.5

These values provide insights into the performance of the regression models in fitting the data.

Conclusion:

From the curve, from figure 2 we can see that mean square error mean apporach is less than mean square error median apporach. We can see that MSE of mean approach is less than Median approach. From fig - 2 mean approach estimator better than median approach because meadian approach estimator is fluctuating much more than mean approach. So mean approach looks quite good and better

Figure 2: Local Polynomial Regression Mean and Median Approach

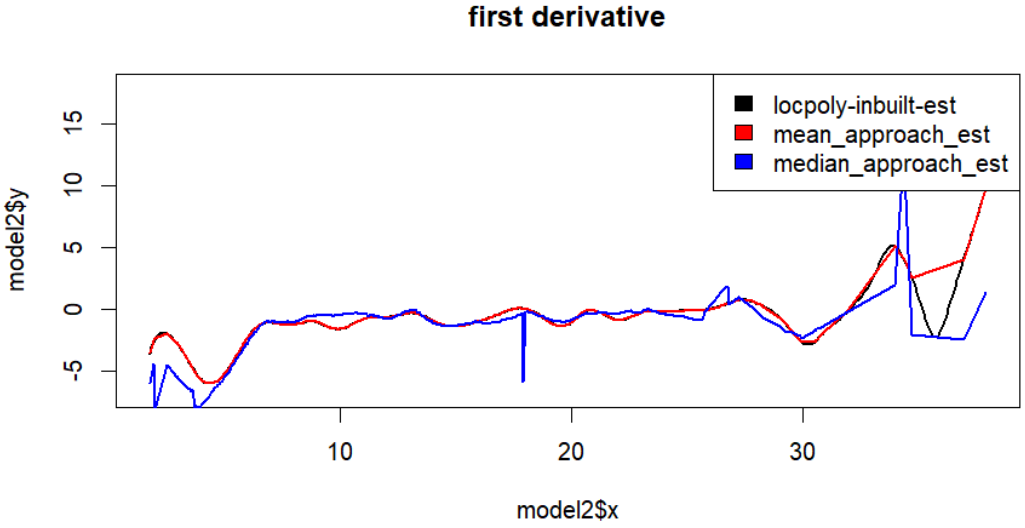

Figure 3: First Derivative

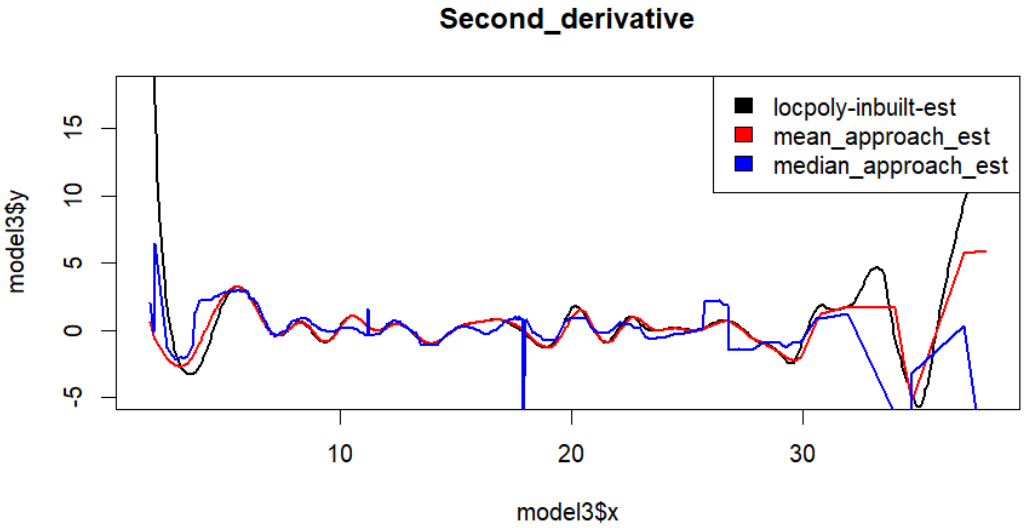

Figure 4: Second Derivative