Assignment 2

Lakshika, Mrunal Dhiwar, Ringan Majumdar Rishiraj Sutar, Sandeep Parmar

2024-02-15

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.

Data Description

Our dataset contains ECG(electro cardiogram) readings of patients. This dataset contains the ECG readings of patients. Each row corresponds to a single complete ECG of a patient. Every single ECG is composed of 140 data points(readings). Columns:-

1. Columns 1-140 contain the ECG data point for a particular patient. These are floating point numbers.
2. The label which shows whether the ECG is normal or abnormal. It is a categorical variable with value either 0 or 1.

We have subsetted 50 observations from each of the categories. Hence, we have 50 observations for two samples each and 140 variables (ensuring p > n). Thus, this gives us a ground to check whether the distributions associated with two samples are independent or not.

We can check whether two samples are independent by using characteristic function. But, in this case the underlying distribution is unknown. Hence, we have used empirical characteristic function.

Methodology

Given two sample X and Y of size $n * p$. We have to check independence of X and Y.

For that we characteristic function approach.

If X and Y independent then

$$\phi_{XY}(t1, t2) = \phi_X(t1)\phi_Y(t2)$$

.

Since we do not have population for characterization so we use empirical characteristic function. Then our equation becomes,

$$\hat{\phi}_{XY}(t_1, t_2) = \hat{\phi}_X(t_1)\hat{\phi}_Y(t_2)$$

where

$$\hat{\phi}_X(t_1) = \frac{1}{n}\sum_{i=1}^{n} e^{i(t_1)'(X_i)}$$

$$X_i$$

is ith obs of X size $p*1$and t_1 is vector of size $p*1$. Similarly we can define $\hat{\phi}_Y(t_2)$. And

$$\hat{\phi}_{XY}(t_1,t_2) = \frac{1}{n}\sum_{i=1}^{n} e^{i(t_1,t_2)'(X_i,Y_i)}$$

Define

$$g(t1,t2) = \hat{\phi}_{XY}(t_1,t_2) - \hat{\phi}_X(t_1)\hat{\phi}_Y(t_2)$$

We calculate $g(t1,t2)$ for 50 different vectors t_1 and t_2 such that $||t_1|| = ||t_2|| = 1$, but projection in p-dimension is different for all vectors.

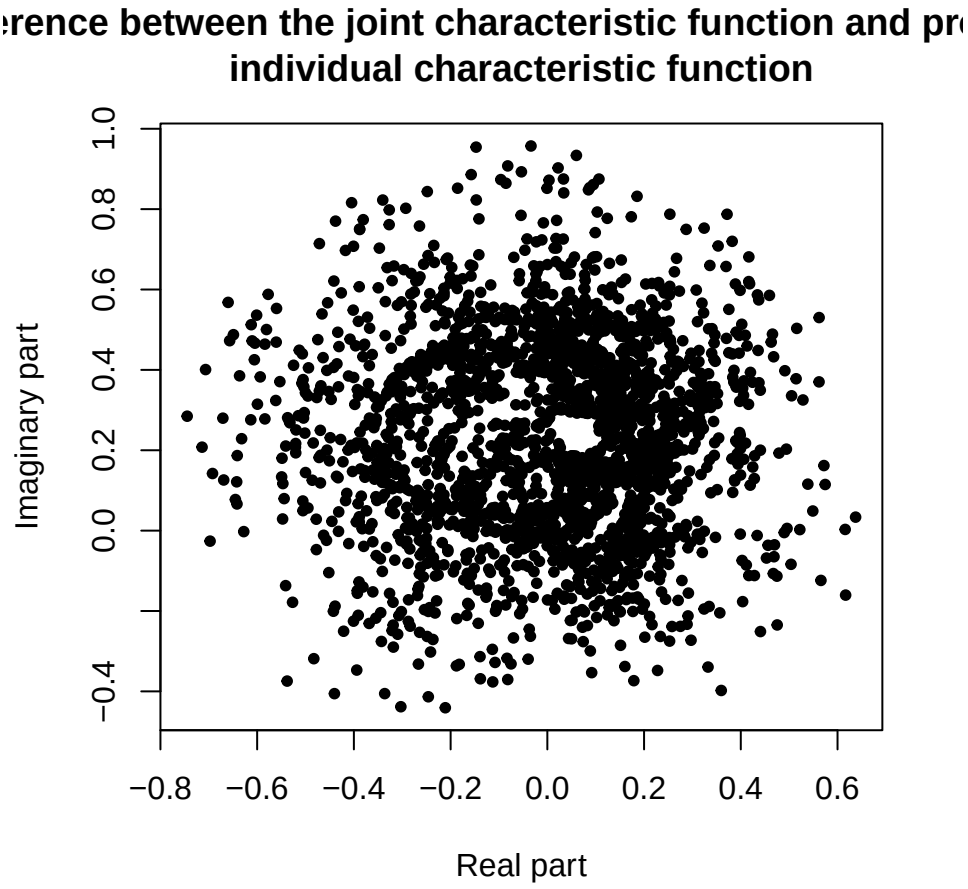
For that we use d-dimensional polar transformation concept. You can check its implementation in `projections(p)` function.

Also for finding joint empirical characteristic function we expand grid of vectors t_1 and t_2, so in total we have 50 x 50 different values of $g(t1,t2)$.

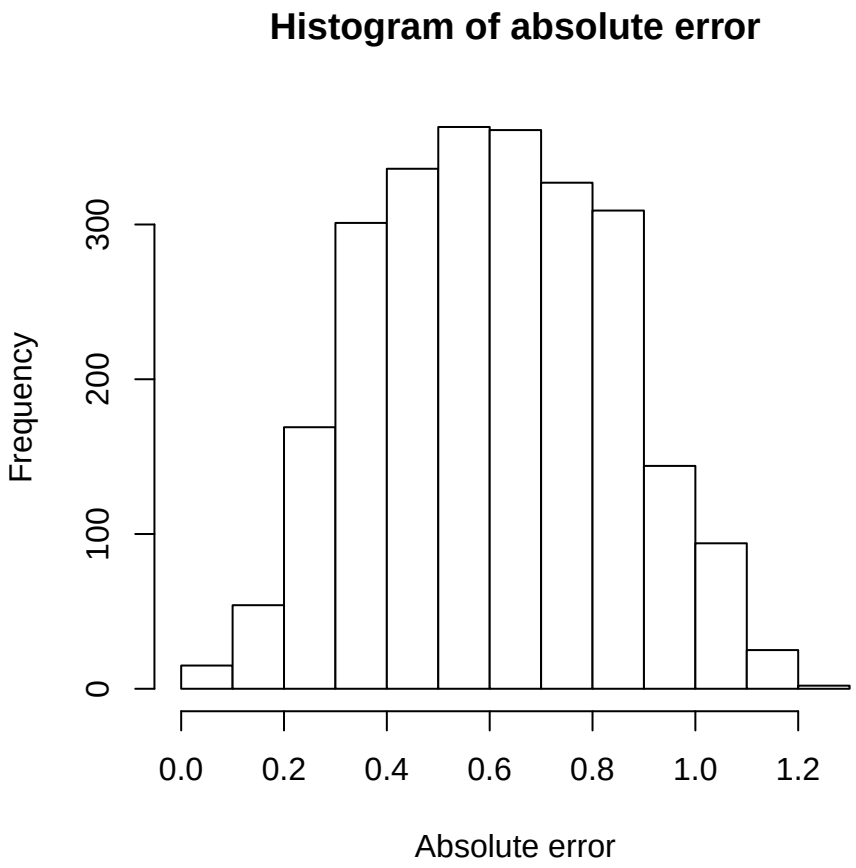
Finally we have 2500 complex number values of $g(t1,t2)$. we take its absolute value and get inference from that.

Plots

Using the above mentioned method, the first plot which we have obtained is the scatter plot of the the complex errors. One can observe that, the points are not clustered around origin of Argand plane. Hence, we can have doubts on the independence of the two samples.



Next, we have plotted the histogram of the absolute values of the errors. Observe that, the values are clustered around (0.2, 0.6) with a mean of 0.443. Hence, this is far from 0. Hence, we might conclude that the samples are not independent.



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.

Data Description

The dataset used in this assignment has been imported from the 'locfit' package in R. It contains information about the Nitrous Oxide (NOx) exhaust emissions from a single cylinder engine. Two predictor variables are E (the engine's equivalence ratio) and C (Compression ratio). In this assignment, we would consider only E (equivalence ratio) as our only predictor.

Non Parametric Regression:

In non-parametric regression, the emphasis is on capturing the underlying structure of the data without imposing strong assumptions about its distribution or form. This makes non-parametric regression particularly useful in situations where the relationship between variables is complex or unknown, or when the data does not adhere to the assumptions of parametric models. Let,$\chi = \{x_1, x_2, \ldots\ldots x_n\}$ denote the predictor variable

and $Y = \{y_1, y_2, ..., y_n\}$ denote the corresponding response variable. If (X_i, Y_i) are iid replications of (X , Y), the model can be written as:

$$Y_i = m(X_i) + \epsilon_i, i = 1(i)n$$

Our target is to estimate m(x), i.e, the regression curve. There are different forms of m(x), namely:

1) location model: $m(x) = \theta$ (constant function)
2) local polynomial regression
3) linear regression: $m(x) = \beta_0 + \beta_1 X$
4) non-linear regression: $m(x) = \beta_0 + sin(\beta)X$

In this assignment, we would only look into the local polynomial estimation.

1) Local Polynomial Mean Estimator:

The Local Mean Estimator is given by:

$$(\hat{m_n}(x_0), \hat{m_n}^{(1)}(x_0), ...\hat{m_n}^{(p)}(x_0)) = \underset{(\theta_0,\theta_1,...,\theta_p)\epsilon R^{p+1}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\{y_i - \theta_0 - \theta_1(x_i - x_0) - \frac{\theta_2}{2!}(x_i - x_0)^2 .... - \frac{\theta_p}{p!}(x_i - x_0)^p\}^2 K(\frac{x_i - x_0}{h_n})$$

where,

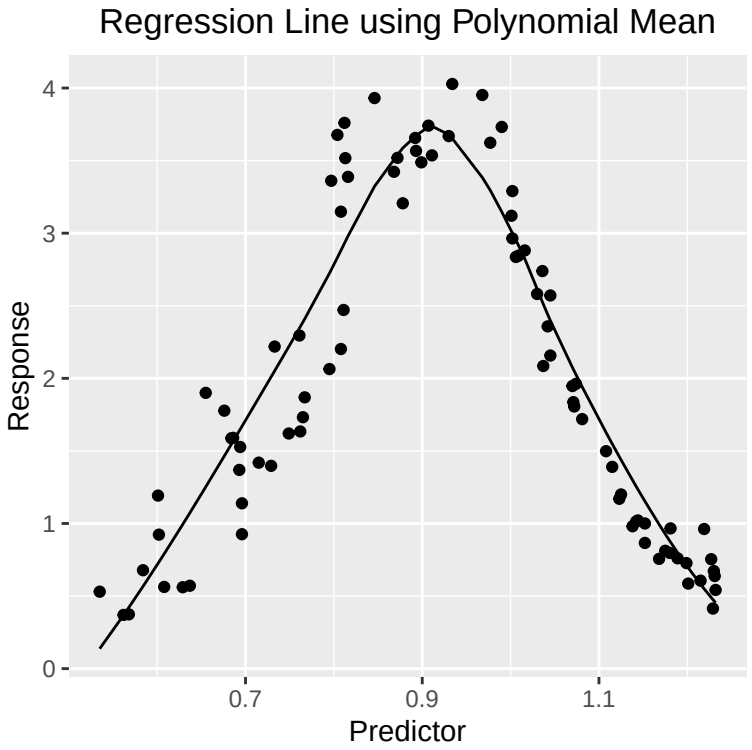
θ_0 estimates $\hat{m_o}(x_o)$, θ_1 estimates $\hat{m_n}^{(1)}(x_0)$, . . θ_p estimates $\hat{m_n}^{(p)}(x_0)$

$K(.)$ denotes the kernel density estimator

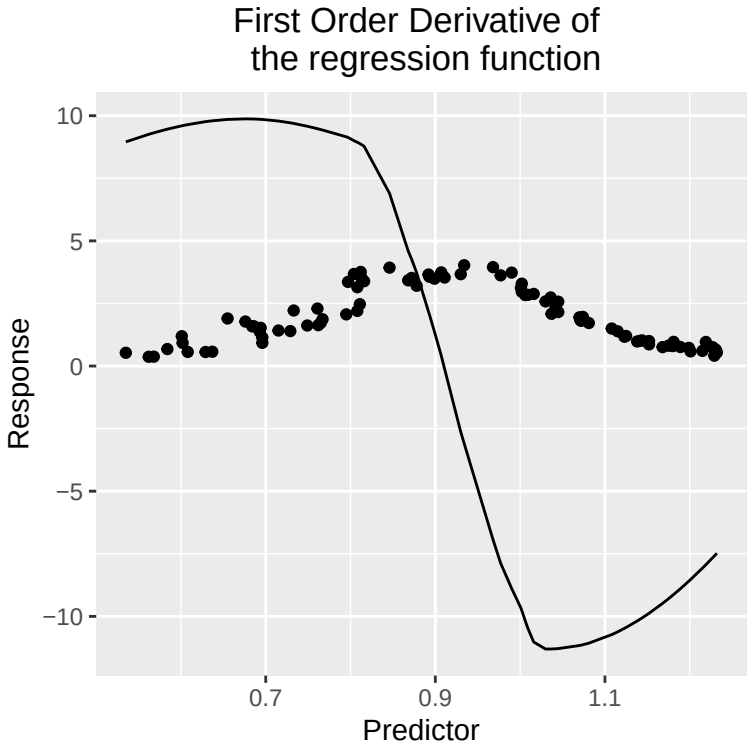
We have used the **locfit.raw** function of locfit package in R to perform the above local polynomial mean estimation.

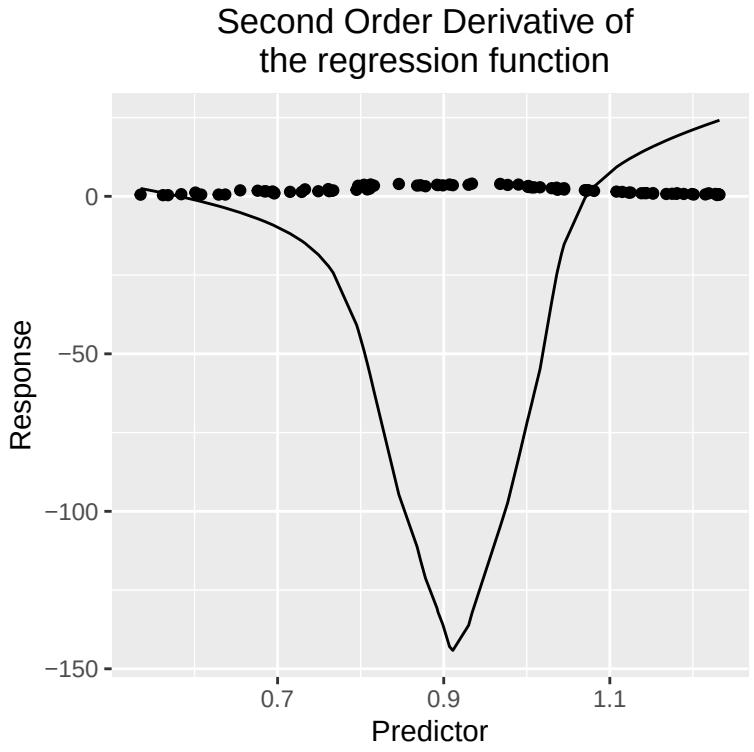
R Code Snippet:

```
library(locfit)
data(ethanol, package="locfit")
data = ethanol
fit <- locfit.raw(data$E , data$NOx , deg = 2 , ev = data$E)
y_hat = predict(fit)
```



Now, we have calculated the first and the second derivatives of the regression function





2) Local Polynomial Median Estimator:

The Local Median Estimator is given by:

$$(\tilde{m_n}(x_0), \tilde{m_n}^{(1)}(x_0), ...\tilde{m_n}^{(p)}(x_0)) = \underset{(\theta_0,\theta_1,...,\theta_p)\epsilon R^{p+1}}{\operatorname{argmin}} \frac{1}{n}\sum_{i=1}^{n}\{y_i - \theta_0 - \theta_1(x_i - x_0) - \frac{\theta_2}{2!}(x_i - x_0)^2 \cdots - \frac{\theta_p}{p!}(x_i - x_0)^p\}^2 K(\frac{x_i - x_0}{h_n})$$

where,

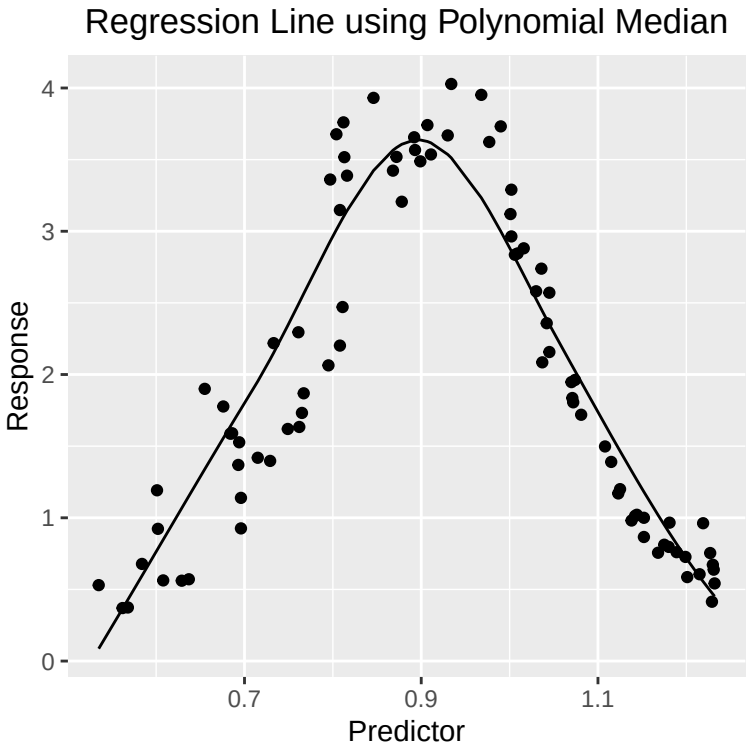
θ_0 estimates $\tilde{m_o}(x_o)$, θ_1 estimates $\tilde{m_n}^{(1)}(x_0)$, . . θ_p estimates $\tilde{m_n}^{(p)}(x_0)$

$K(.)$ denotes the kernel density estimator

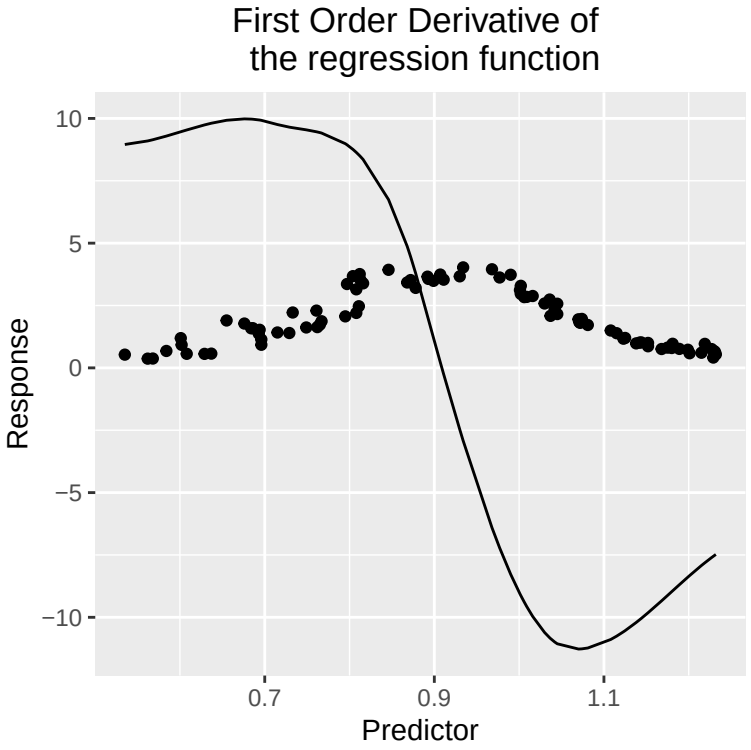
We have used the **locfit.roburst** function of locfit package in R to perform the above local polynomial median estimation.

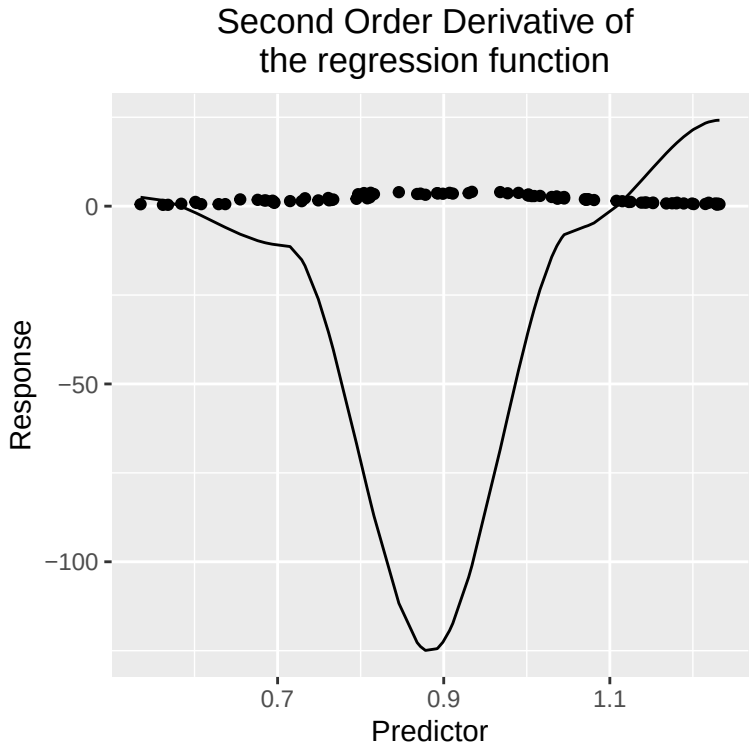
R Code Snippet:

```
library(locfit)
data(ethanol, package="locfit")
data = ethanol
fit <- locfit.robust(x = data$E , y = data$NOx)
y_hat = predict(fit , data$E)
```



Now, we have calculated the first and the second derivatives of the regression function





Model Evaluation:

Now, to compare the performance of the two models, we would calculate the mean square error (MSE) of both the models.

The formula for Mean Square Error is given by:

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2$$

where Y_i = actual value and $\hat{Y}_i$ = predicted value

- MSE for local polynomial mean = 0.3547
- MSE for local polynomial median = 0.3836

Given that the Mean Squared Error (MSE) for the first model is marginally lesser than that of the second model, it can be inferred that the local polynomial mean regression demonstrates superior performance compared to the local polynomial median regression.

References

1. *J.L.O. Cabrera. locpol: Kernel local polynomial regression, 2009. URL http://CRAN.R-project.org/package=locpol. R package version 0.4-0.*
2. *Derivative Estimation using Local Polynomial Fitting: De Brabanter, Kris; De Brabanter, Joseph; Gijbels, Irène; De Moor, Bart. Journal of Machine Learning Research ; 2013; Vol. 14.*