STA-6413 Final Project Report

Group 1 - *Kamaniya Chatakondu, Yahui Peng, Ummugulsum Arslan*

1. Goal and Conjecture

Ninety-nine randomly generated observations with one dependent variable and one independent variable are given as the data of interest. Our goal is to find the best nonparametric regression line.

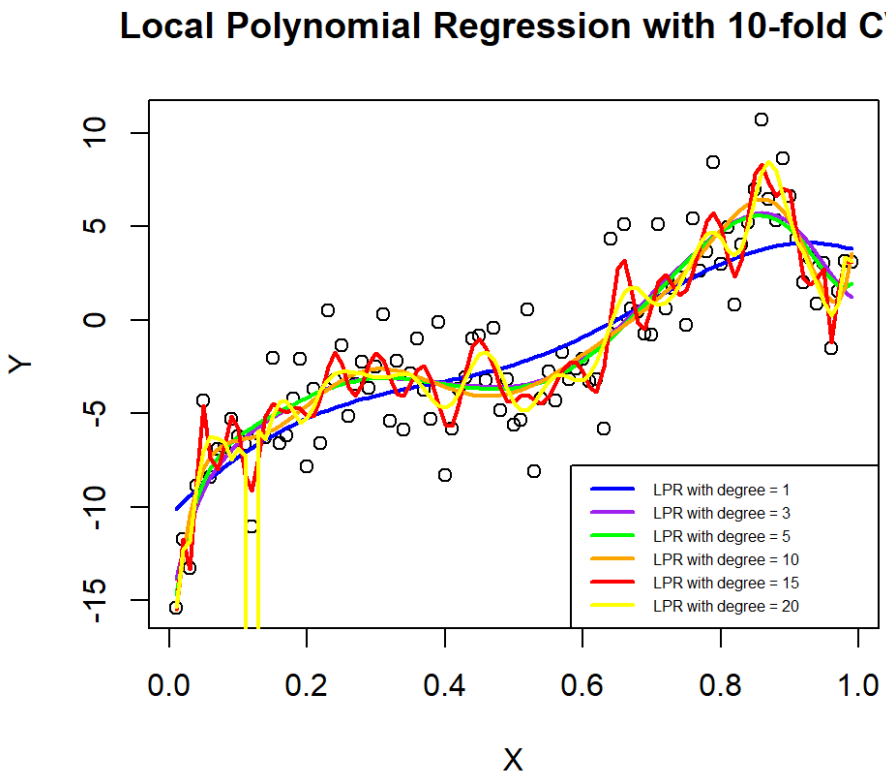
2. Modeling and Performance Evaluation

Local polynomial regression and cubic spline methods are chosen to fit the data. Additionally, ten-fold cross-validation is used to evaluate the performance of the models.

2.1 Local polynomial regression (LPR)

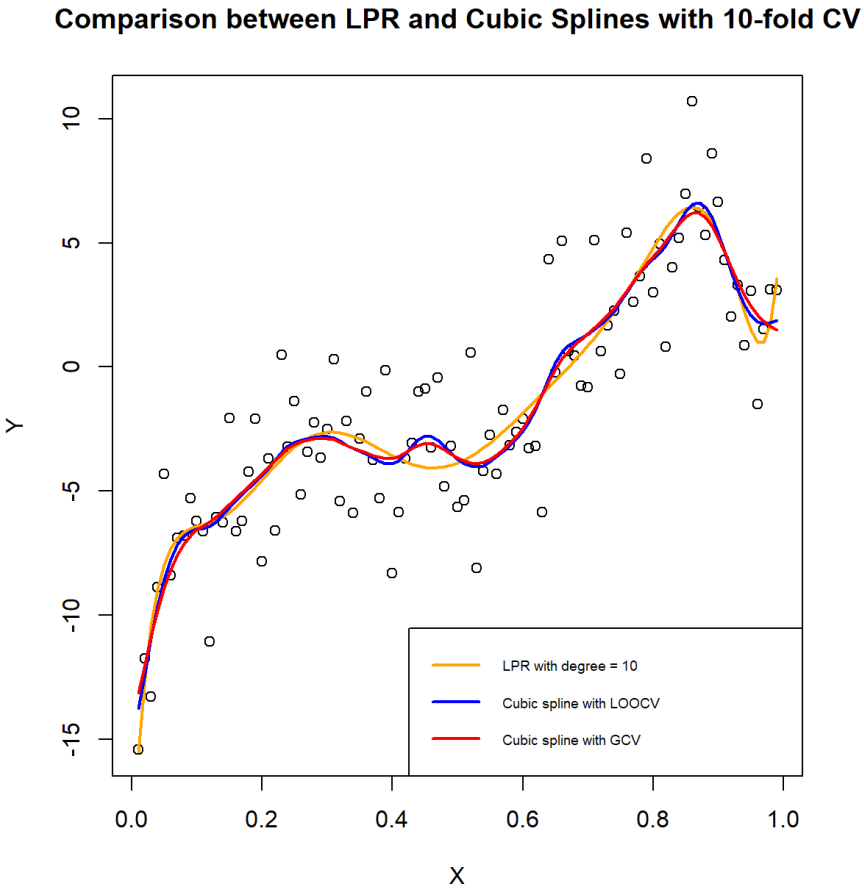
LPR is a nonparametric technique for smoothing scatter points and modeling functions using only points in some neighborhood (bandwidth) of each point. LPR has good performance on the boundary and is superior to all other linear smoothers in a minimax way. The performance of the predicted function is dependent on the selection of the kernel function, the size of neighborhood, and the degrees of freedom of the polynomial fit.

The *sharpData* package in R is used to perform cross-validation of the LPR. We tune the degree of the model with 1, 3, 5, 10, 15, and 20 under the Gaussian kernel function. The mean squared errors (MSEs) achieved are 6.97, 5.24, 4.92, 2.81 and 2258.21, accordingly. It's observed that degrees of 15 and 10 have the lowest and second lowest MSE, respectively. Additionally, the fitted LPR curves are plotted below. The curve of degree = 10 is smooth and follows the trend of the data points well. However, the curve of degree = 15 is wiggly-shaped, suggesting potent overfitting issues. Therefore, 10 is selected as the best LPR degree with bandwidth at 0.66.



2.2 Cubic Spline

Cubic spline possesses advantages in terms of simplicity of calculation, numerical stability and smoothness of the interpolated curve. Hence, we fit cubic splines to the data besides LPR. The "cv" argument in *smooth.spline*() function is set to TRUE and FALSE to compare the result with LOOCV and generalized cross-validation (GCV). The MSEs obtained for LOOCV and GCV are 4.51 and 4.81, respectively, which are both slightly lower than that of LPR with degree = 10 (4.92). Moreover, the fitted LPR and cubic spline curves are plotted below. All curves follow the trend of data points well and do not seem to be wiggly enough for an overfitting issue. Therefore, cubic splines perform slightly better than LPR with degree = 10, and LOOCV offers the best model performance with the lowest MSE of 4.51.



3. Conclusion

LPR and cubic spline methods are applied to fit the data of interest in nonparametric sense. Comparing the MSE obtained with 10-fold cross validation, we conclude that cubic spline with LOOCV offers the best model performance while avoiding an overfitting issue.