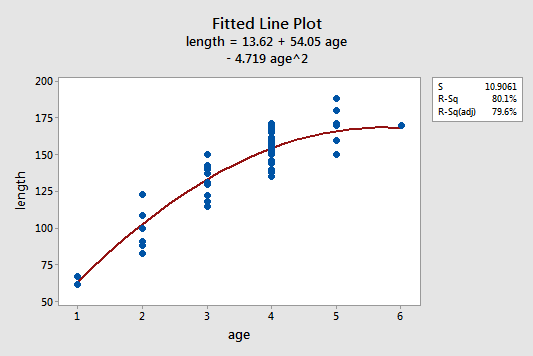
**Robust Polynomial Regression Using Kernelized Spatial Depth, and an Application**

In 1981, 78 bluegills were randomly sampled from Lake Mary in Minnesota. Researchers measured and recorded the following data:

Response (y): length (in mm) of the fish; and Potential predictor (x): age (in years) of the fish

The researchers were primarily interested in learning how the length of a bluegill fish is related to it age. A scatter plot of the data suggests that there is positive trend in the data. This is not surprising as the age of bluegill fish increases, the length of the fish tends to increase. The trend, however, does not appear to be quite linear. It appears as if the relationship is slightly curved. The estimated quadratic regression function looks as follows:



We are interested in the following questions:

How to account for the variability in the length for each age? Will the results improve if we do a weighted average for each age?

The challenge is to draw the best fit curve over a data which contains some outliers. The usual mean is severely affected by the presence of outlying observations, namely, the outliers. The general method of least squares will yield misleading results, and it will no longer express the precise trend that data is following. So, we plan to work with the median and the weighted mean to design a robust regression method to handle such regression problems.

We propose a novel framework based on the notion of statistical depths (see Liu, Parelius and Singh, 1999). Most of the current depth methods fail scale with the dimensionality of the input spaces. More specifically, they have an order of complexity O(n^{d/2}), where n is the sample size and d is dimension of the input space. We will use a new depth function, the kernelized spatial depth (KSPD) (see Chen, Dang, Peng and Bart Jr., 2009), which defines the spatial depth in a feature space induced by a positive definite kernel. The KSPD of an observation can be evaluated directly from the data set with a computational complexity O(n^2) which is independent on d. Moreover, KSPD identifies outliers for more complex data (e.g., data with non-linear trends).

We plan to design appropriate simulation schemes to thoroughly analyze, and establish reliability of our regression method using KSPD. The proposed method will be used to analyze the real data, and draw relevant inference.

The given research proposal shall be rigorously completed under the supervision of Prof. Subhajit Dutta in the time period of 8 weeks.

***References:***

1. Chen, Yixin, Dang, Xin, Peng, Hanxiang and Bart Jr., Henry L. (2009) Outlier Detection with the Kernelized Spatial Depth Function. IEEE Transactions on Pattern Analysis and Machine Intelligence, 288-305.

2. Liu, Regina, Y. Parelius, Jesse M. and Singh, Kesar (1999) Multivariate analysis by data depth: descriptive statistics, graphics and inference, (with discussion and a rejoinder by Liu and Singh). The Annals of Statistics, 27, 3, 783-858.

3. Data: https://onlinecourses.science.psu.edu/stat501/node/325.