Analysis of Seismic Timing

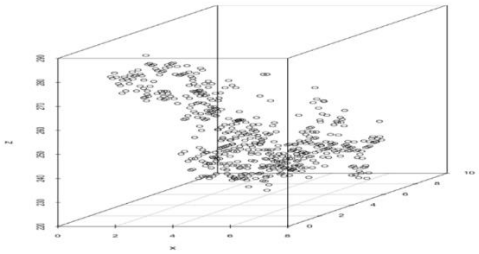
Data 583: Advanced Predictive Modelling

Himabindu Joopally, Wei Tang, Harpreet Singh

1. Introduction

Seismic timing data: The z variable of this data set corresponds to seismic timings measured by geophones dropped into ditches dug along transects following the (x, y) coordinates. The timings are related to depth of a substratum. The shape of this surface is of importance in oil exploration. This substratum represents an ancient riverbed (river bottom) in central Alberta.

3D Visualization of the Data:

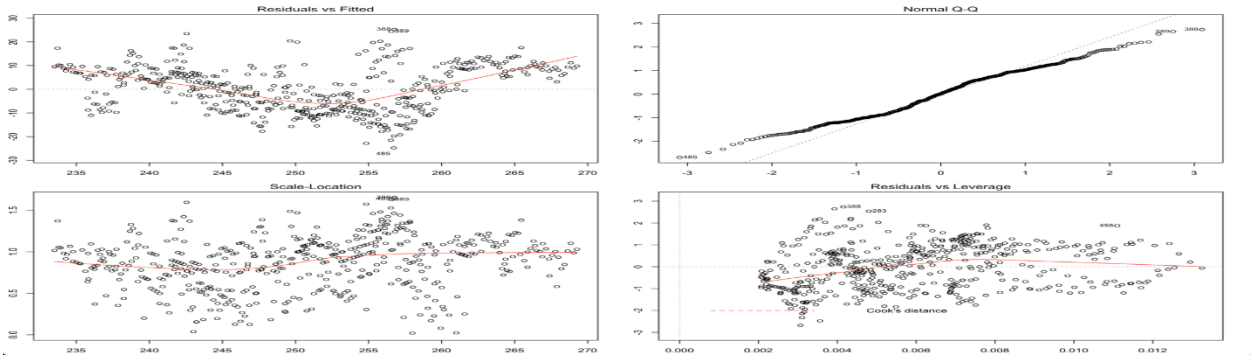


2. Models Considered

Multiple Linear Regression:

The first model is the simplest model which is linear regression. We assume there is a linear relation between response variable z, and variables x and y.

Below are the residual plots returned by R:

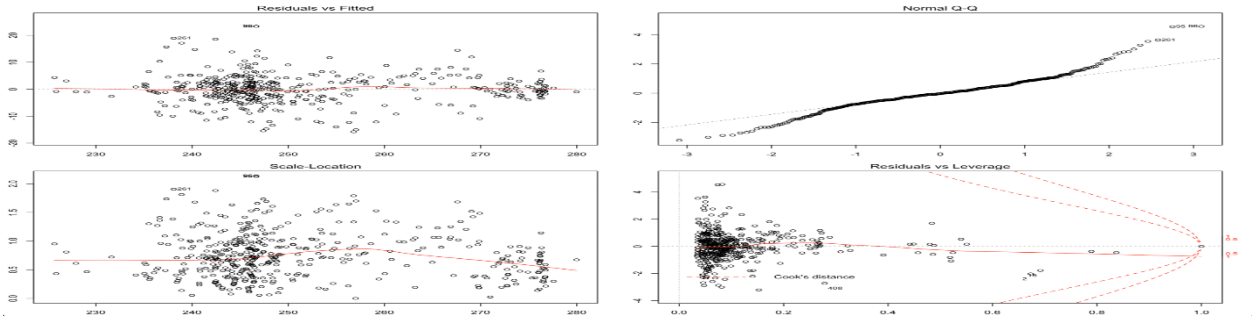


As we can see from the first plot Residual vs Fitted, it suggests a non-linear pattern in the data. Low adjusted R-squared and high AIC also support that the model has high bias. Linear regression model is not a good fit to this dataset.

Bivariate Spline Regression:

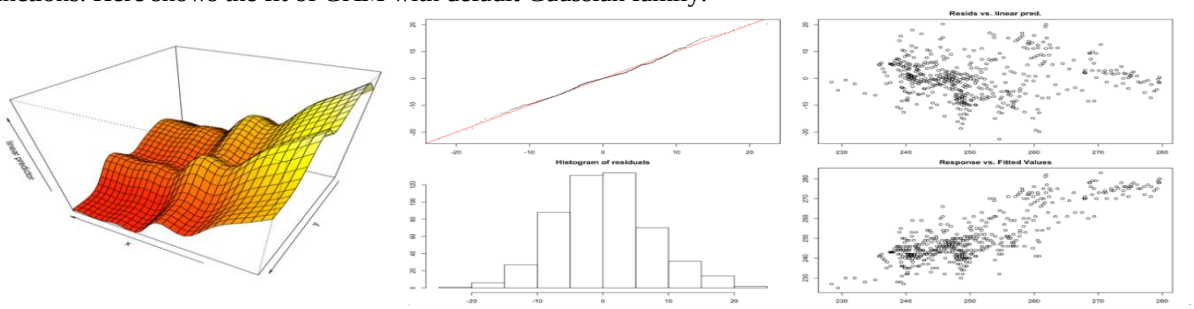
The second model we used is B-spline with 5 equally spaced knots. With this model, we assumed two relations, with and without interaction between variables x and y.

Model without interaction between x and y gives a better fit, with a higher adjusted R-squared and a lower AIC than linear regression. However, if we consider bivariate splines with interaction term x*y, the adjusted R-squared increases to 0.823, and AIC value decreases to 3173.01. Moreover, plot Residual vs Fitted shows no clear pattern, which matches out randomness assumption. Bivariate spline regression with interaction seems to be a good fit till now.



Generalized Additive Model with Gaussian Family:

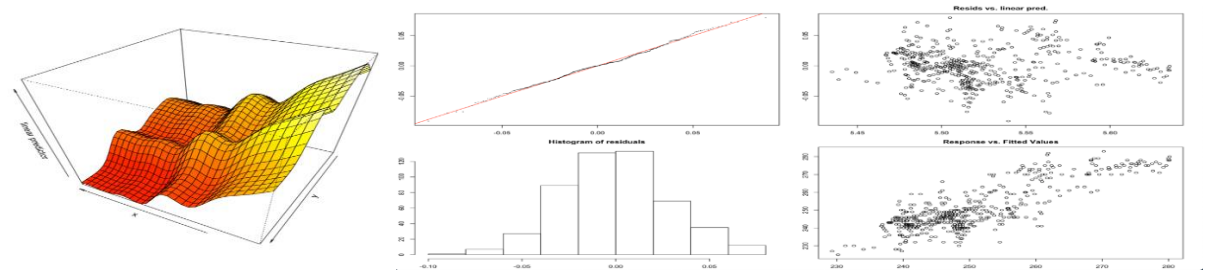
With generalized additive models (GAM), instead of linear relation between predictors, we use the sum of smoothing functions. Here shows the fit of GAM with default Gaussian family.



Adjusted R-squared and AIC are close to bivariate spline regression without interaction, which indicate a poor fit. In addition, the last plot on right Responses vs. Fitted Values has dispersed points around the straight line. The response variable z is always positive, a generalized additive model, using a nonnegative distribution for the response variable might be more realistic.

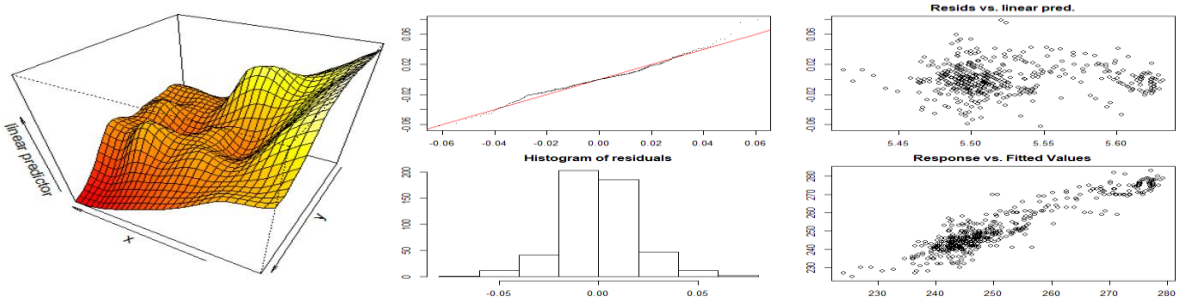
Generalized Additive Model with Gamma Family:

In this model fit, we tried using Gamma family. Not surprisingly, it returned a similar result as previous model.



Generalized Additive Model with Thin-plate Spline:

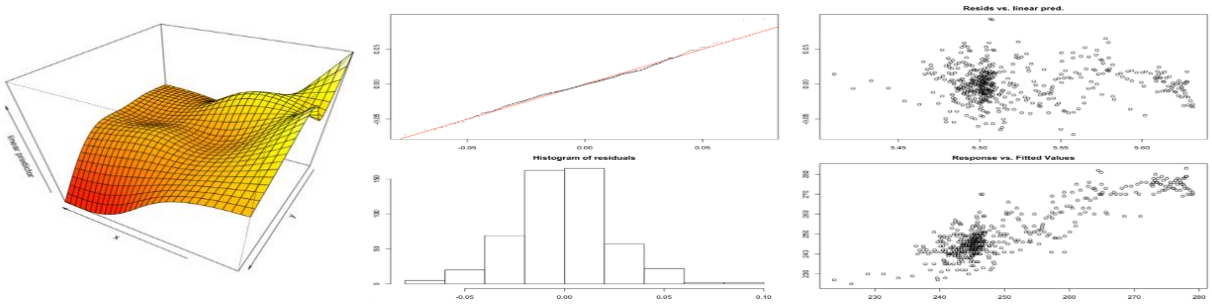
A thin-plate spline is a higher-dimensional version of a smoothing-spline.



The fit looks better with adjusted R-squared reached 0.849 and AIC lowered to 3078. The Response vs. Fitted Values plot also looks much less dispersed around the straight line than before.

Generalized Additive Model with Tensor-product Spline:

A tensor-product spline can be computationally more efficient than a thin-plate spline. In this case, x and y are in the same units and we expect the same wiggles in both variables. So this model will be similar to thin-plate spline model.



The fit looks smoother than thin-plate spline, while adjusted R-squared decreased and AIC increased. Furthermore, the Response vs. Fitted Values plot looks a little bit more dispersed than thin-plate spline model.

Summary Table:

Model	Adjusted R-squared	AIC	Deviance Explained
Multiple Linear Regression	0.469	3676.541	47.16%
Bivariate Spline Regression (w/o interaction)	0.645	3487.768	65.63%
Bivariate Spline Regression (with interaction)	0.823	3173.010	84.17%
GAM Gaussian family	0.676	3442.686	68.60%
GAM Gamma family	0.678	3432.615	65.40%
GAM Thin Plate Spline	0.849	3078.495	85.60%
GAM Tensor Product Spline	0.764	3286.900	76.90%

3. Conclusion

Generalized additive model with thin-plate spline has the best AIC, adjusted R-squared and deviance explained values among all models considered. However, we don't have test data, and original dataset is too small to be split into train and test, it is hard to decide whether the model is overfitting or not.

The following is the contour plot of GAM with thin-plate spline. The shade of the colour gets darker at top right corner, and lighter at left side. The larger (x, y) coordinates, the smaller z, and vice versa.