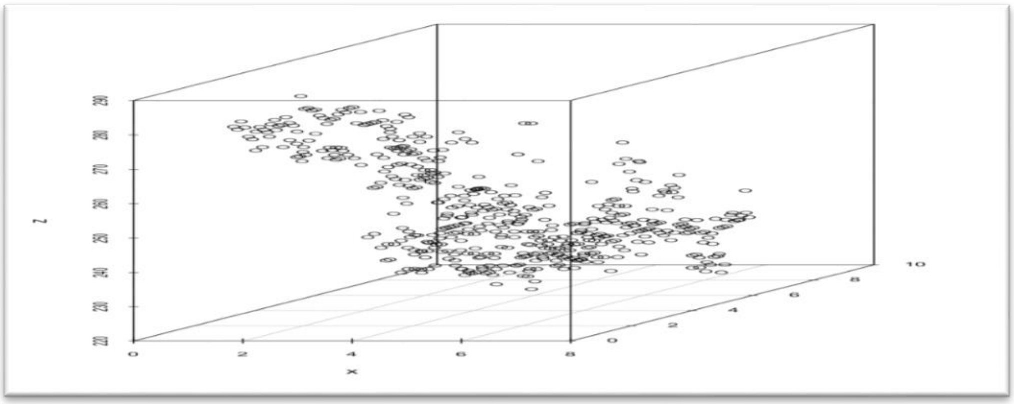
DATA – 583 Project

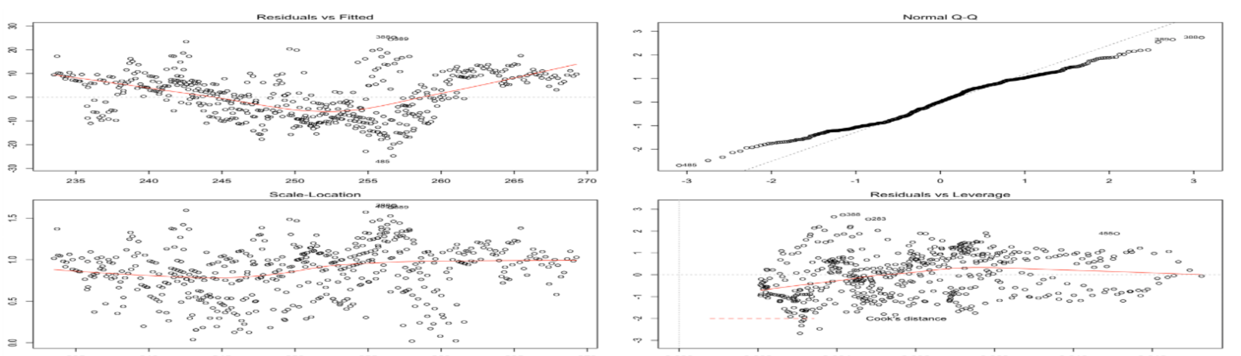
Seismic timing data - The z variable of this data set corresponds to seismic timings measured by geophones dropped into ditches dug along transects following the (x,y) coordinates. The timings are related to depth of a substratum. The shape of this surface is of importance in oil exploration. This substratum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

3D Visualization of the Data:



Models Considered:

Multiple Linear Regression:

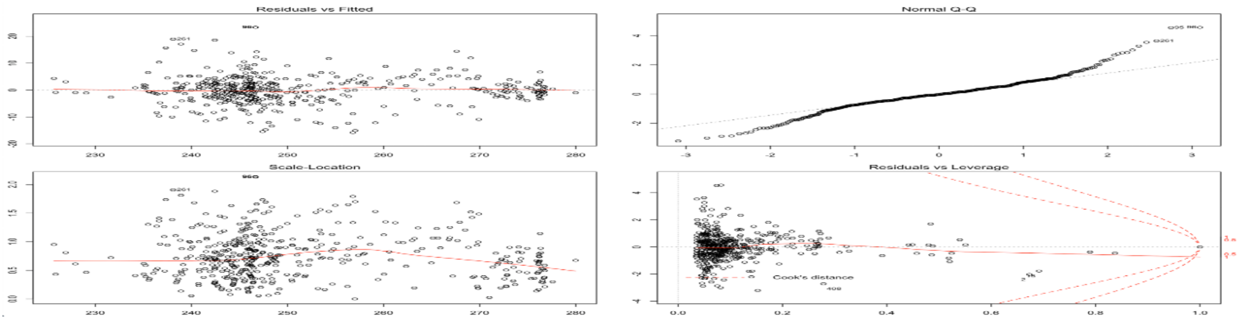


The first model is the simplest model which is linear regression. The first plot suggests some non-linear pattern in the data.

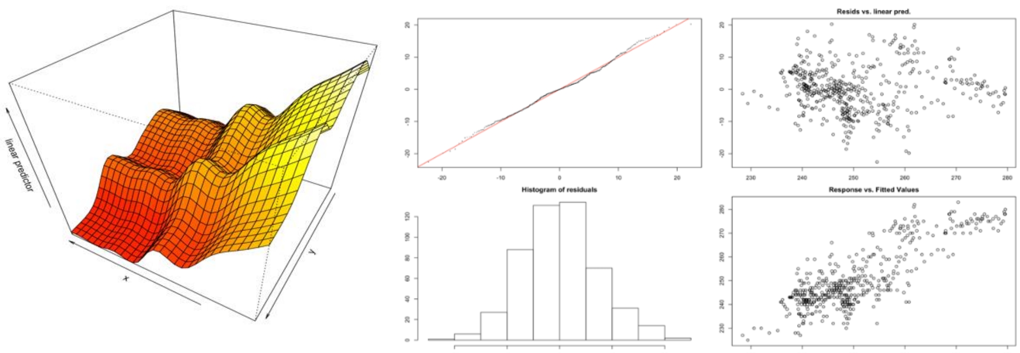
- Adjusted R-squared value is 0.4695 which is too low and indicates that the model has high bias. More complicated models like splines should be considered.
- AIC value is 3676.541
- Deviance explained = 47.16%

Bivariate spline regression with 5 equally spaced knots (With-out and with interaction term)

Adjusted R-squared is 0.645 which is better than MLR but still not very good. AIC value is 3487.768. Deviance explained = 65.63%. If we consider bivariate splines with interaction term x*y, the adjusted R-squared increases to 0.8227. AIC value is 3173.011, Deviance explained = 84.17%.



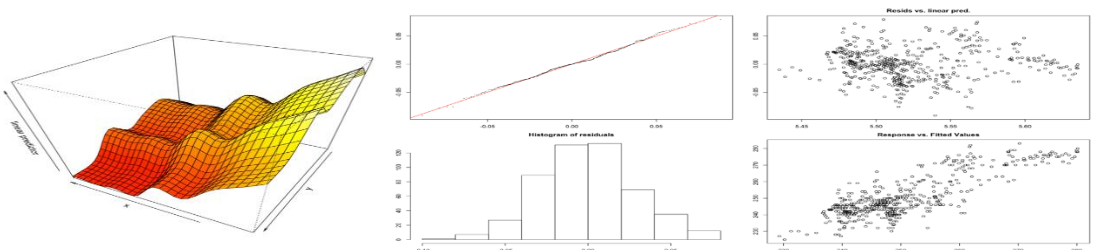
Generalized additive models with default Gaussian family



Adjusted R-squared is 0.676, and deviance explained = 68.6%, AIC = 3442.686(Not so good). Visualizing the goodness of fit (plot on the left) Plots on the right: All of them look good, but the Response vs. Fitted Values plot looks very dispersed around the straight line. The response variable z is always positive, so a generalized additive model, using a nonnegative distribution for the response variable might be more realistic. So, we try using Gamma family.

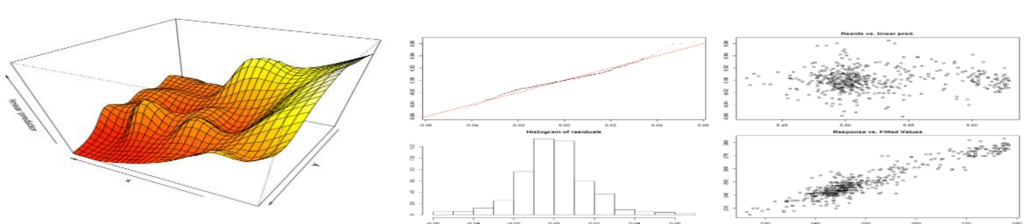
Generalized additive models with default Gamma family

Adjusted R-squared = 0.678, Deviance explained = 68.4%, AIC = 3432.615.All the plots and the values look very similar to the previous model with gaussian family.



5 5 5 **5)**

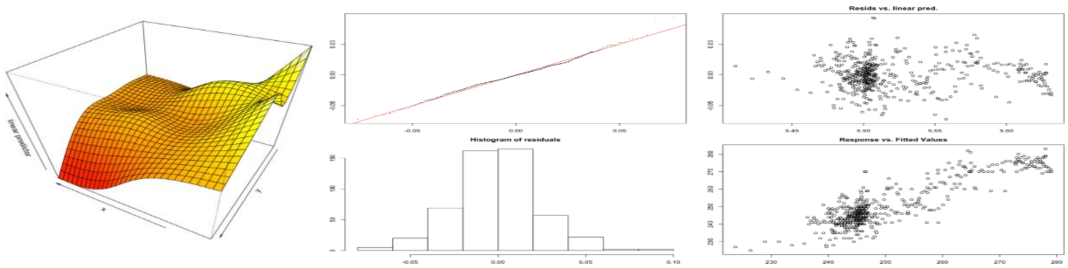
Generalized additive models with thin-plate splines



A thin-plate spline is a higher-dimensional version of a smoothing-spline. The fit looks better. The Response vs. Fitted Values plot looks lot less dispersed around the straight line, which indicates better fit. Adjusted R-squared = 0.87, Deviance explained = 87.9%, AIC = 3012.056

Generalized additive models with tensor-product splines

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 may be similar to thin-plate splines.



The fit looks smoother than thin-plate spline. But the Response vs. Fitted Values plot looks a little bit more dispersed. Adjusted R-squared = 0.764, Deviance explained = 76.9%, AIC = 3286.908

Model Evaluation:

Model	Adjusted R-squared	AIC value	Deviance explained
Multiple Linear Regression	0.4695	3676.541	47.16%
Bivariate Spline regression	0.645	3487.768	65.63%
GAM Gaussian family	0.676	3442.686	68.6%,
GAM Gamma family	0.678	3432.615	65.40%
GAM Thin plate splines	0.87	3012.056	87.90%
GAM Tensor product Spline	0.764	3286.90	76.90%

Selected model:

Generalized additive models with thin-plate splines has the best AIC, Adjusted R squared and Deviance values among the models considered. The fit also looks good and reasonably smooth.