**DATA 583 Project**

**About the data:**

**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 particular substratum. The shape of this surface is of importance in oil exploration. This particular substratum represents an ancient riverbed (river bottom) in central Alberta.

**3D Visualization of data:**

**A close up of a map

Description automatically generated**

**Models considered:**

1. **Multiple linear regression**

A close up of a map

Description automatically generated

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%

1. **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%

A close up of a map

Description automatically generated

1. **Generalized additive models with default Gaussian family**

Adjusted R-squared is 0.676, and deviance explained = 68.6%, AIC = 3442.686

Not so good.

A close up of a logo

Description automatically generatedA close up of a map

Description automatically generated

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.

1. **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.

A close up of a logo

Description automatically generatedA close up of a map

Description automatically generated

1. **Generalized additive models with thin-plate splines**

**A close up of a logo

Description automatically generatedA close up of a map

Description automatically generated**

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

1. **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 wiggliness in both variables, so this model may be similar to thin-plate splines.

**A picture containing text

Description automatically generatedA close up of a map

Description automatically generated**

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

**Add a table for model comparison**

**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.

**Contour Map:**

A picture containing text

Description automatically generated