DATA 583 Project

Bruce Pei, Liza Wood, Mohsen Zardadi

09/04/2019

Exploring the seismic timing data

Before we start modelling and determine which model is the best fit, it is a good practice to visualize the data to get a sense of what we are looking at. We did that, but it doesn't fit in a 3-page report.

We are clearly dealing with non-linear, 3-dimensional data. `z` is the response while `x` and `y` are our variables. Since `x` and `y` are coordinates, there is no interaction between those two variables.

To determine which model is the best fit, we looked at multiple linear regression, bivariate spline regression with equally spaced knots, generalized additive models with normal and gamma families, thin-plate splines and tensor-product spline. To compare the models, we looked at AIC across models, along with diagnostics specific to each model. We looked at the plots for all models, but are only including the plots for the best model in this report.

Muliple Linear Regression

We started with using Multiple Linear Regression which gave the following output for R^2 and AIC:

```
## [1] "R-squared for multiple linear regression: 0.471620164063547"
```

```
## [1] "AIC for multiple linear regression is: 3676.54070379692"
```

The large AIC and low R^2 don't support multiple linear regression as being an appropriate model. Acknowledging that the data isn't linear, we even tried a regression on x^2 and y^2 with similar results.

Comparison of Models Using `gam()` Function

Using the `gam()` function we created the following models: - bivariate spline regression with 20 equally spaced knots - general additive model (GAM) with normal family - general additive model (GAM) with gamma family - thin plate spline - tensor product spline

To compare the above models we looked at percent deviance explained, GCV, and AIC. The best of the above models would be the one with the largest percent deviance explained, the smallest GCV and the smallest AIC. For the last two models we did not specify the number of knots and let `gam()` calculate it from the null space dimension.

Comparing the percent deviance explained:

```
## [1] "Deviance Explained by Model:"
```

```
## [1] "Bivariate Spline Regression (seis_SR_ESK_N): 0.509842979195326"
```

```
## [1] "GAM with normal family (seis_GAM_N): 0.686425778165463"
```

```
## [1] "GAM with gamma family (seis_GAM_G): 0.68441515399619"
```

```
## [1] "Thin-Plate Spline (seis_TPSL: 0.796404146391926"
```

```
## [1] "Tensor Product Spline (seis_TP: 0.768856617687805"
```

Comparing the GCV:

```
## [1] "GCV by Model:"
```

```
## [1] "Bivariate Spline Regression (seis_SR_ESK_N): 82.7960722587088"
## [1] "GAM with normal family (seis_GAM_N): 54.0616119221376"
## [1] "GAM with gamma family (seis_GAM_G): 0.000845653237925553"
## [1] "Thin Plate Spline (seis_TPSL): 0.000570534509958431"
## [1] "Tensor Product Spline (seis_TP): 0.000633960411866849"
```

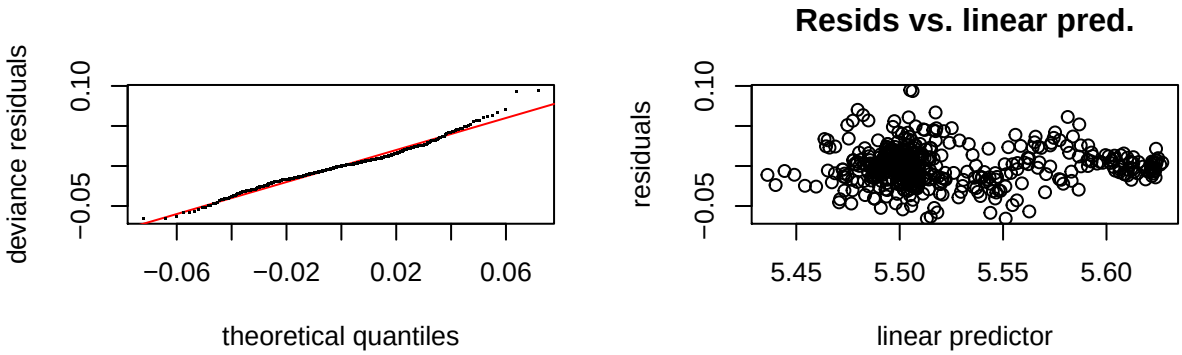
Comparing the AIC:

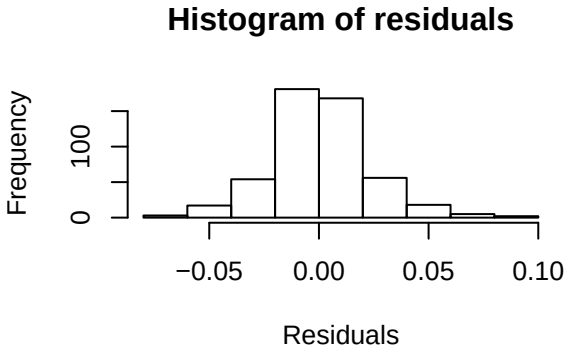
```
##                       df      AIC
## seis_MLR        4.00000 3676.541
## seis_SR_ESK_N  13.56587 3657.827
## seis_GAM_N     18.56115 3442.686
## seis_GAM_G     18.61607 3432.615
## seis_TPSL      29.37923 3233.220
## seis_TP        24.24151 3286.908
```

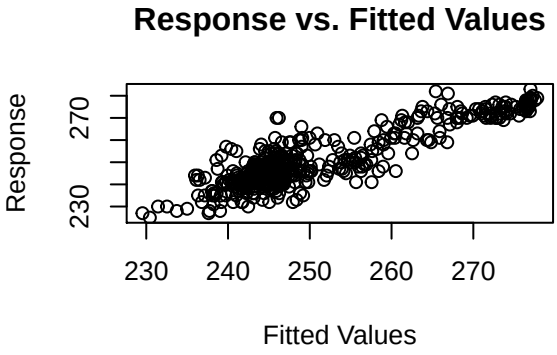
Since `z` is always positive, using a non-negative gamma distribution for the response variable is more realistic. As a result, both thin-plate spline and tensor-product were run with gamma families after we saw the improvement with GAM using gamma families.

Based on the criteria, thin-plate spline is the best model. It has the largest percent deviance explained and the smallest GCV and AIC.

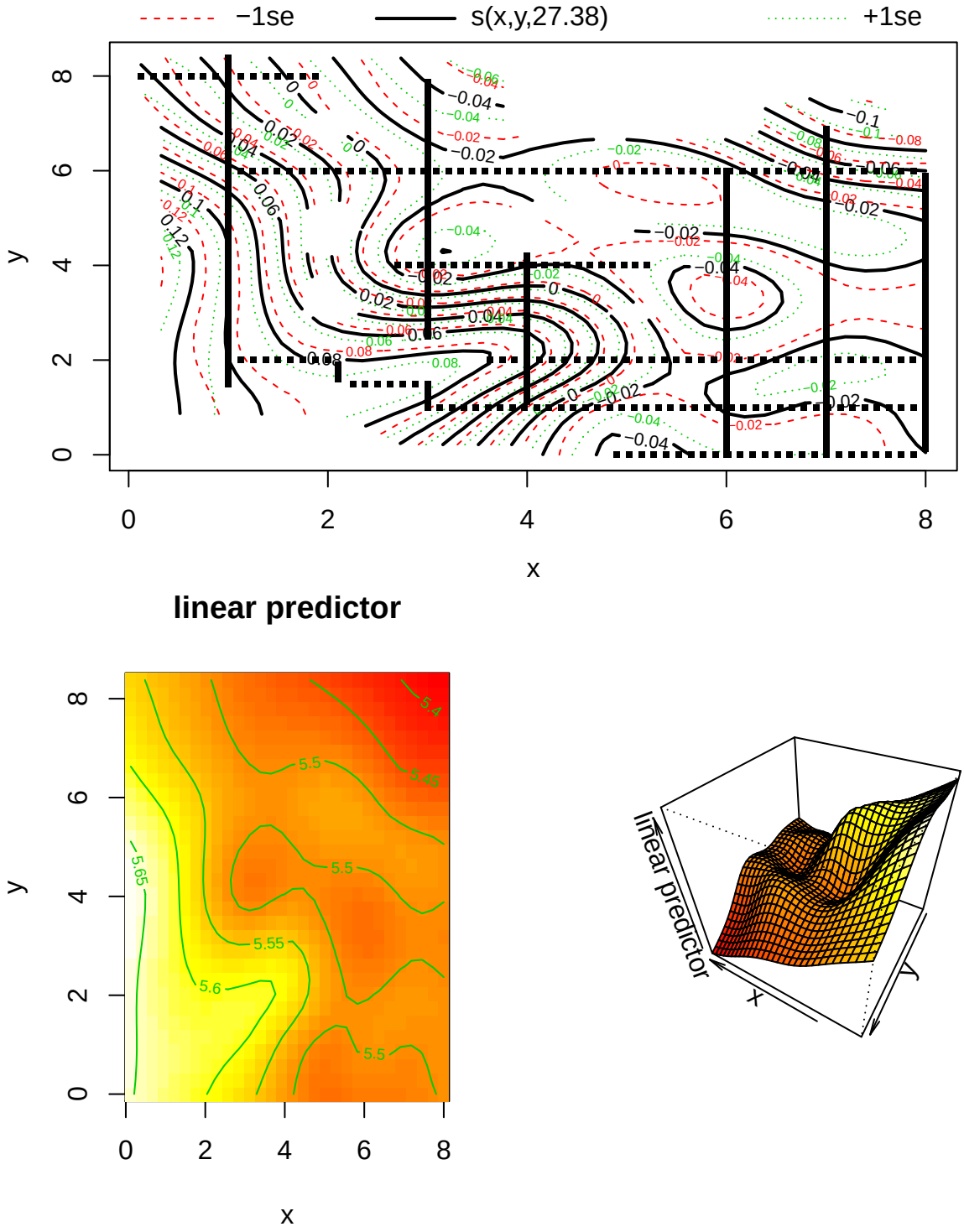
We checked the diagnostic charts for the model and they also supported that this was a reasonable model.







One of the benefits of thin-plate spline and (tensor-product) models is the x,y contour plot, which is a representation of 3D data that a geologist could use. Let's see the plots:



The above plots could be given to a geologist to give a sense of where the different geological features are located.