Seismic Timing Project Report

Multiple Linear Regression Model:

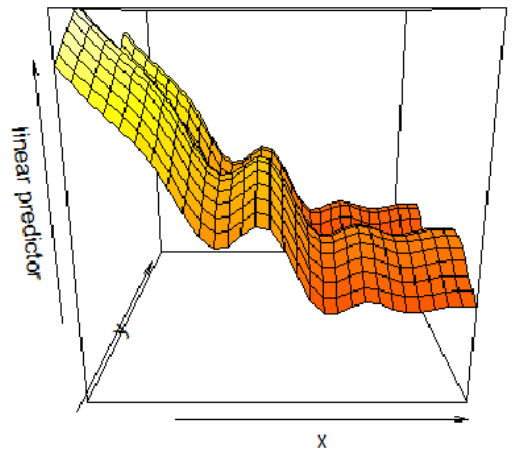
```
seis.lm <- lm(z~y+x, data=seis)
```

The multiple linear regression aims to model the linear relationship between the two explanatory variables x and y (coordinates of the transects) and the response variable z (seismic timing). From the output of the model, the adjusted R-squared value is 0.4695 which means the model explains only 46.95% variance among the data. Running AIC function on this model, it gives an AIC score of 3676.541. The smaller the AIC score, the better the fit.

Bivariate Spline Regression Model (with equally spaced knots):

```
number.knots <- 10
spacings_x <- seq(from=min(seis$x),to=max(seis$x),length=number.knots+2)[2:(number.knots+1)]
spacings_y <- seq(from=min(seis$y),to=max(seis$y),length=number.knots+2)[2:(number.knots+1)]
seis.sr <- gam(z ~ s(x) + s(y),knots=list(spacings_x,spacings_y), data = seis)
```

Generalized additive model package is chosen here for fitting a smooth-spline regression. Specifically, the main fitting function used is gam() with default smoothing functions s() to fit penalized cubic regression splines to the data. From the output of the model, the adjusted R-squared value is 0.676, the ACI score is 3442.686, and the GCV for this gam model is 54.062. The absolute value of GCV score can be interpreted as an estimation of the lack of fit of the model. The smaller the GCV value, the better the model.
The visualization of the model fit is shown below:

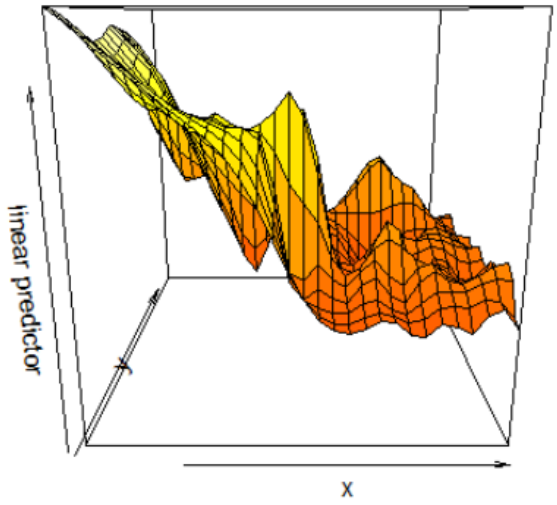


As can be seen in the plot, the model fits the data smoothly, but the degree of the fit seems insufficient. Also according to the R^2 value, only 67.6% variance can be captured by this model.

Thin-Plate Spline Model:

```
seis.tpsl <- gam(z~s(x,y,k=140), data=seis)
```

A thin-plate spline is a higher-dimensional version of a smoothing-spline with fewer knots (and this is controlled with the parameter k). To select the optimal value for k, gam.check() function is applied on the created thin-plate spline model. This will give an output with a p-value which may indicate that k is too low if the given p-value is low (especially if edf value is close to k). In this case, the p-value becomes large when k = 140. Using this parameter, this model gives an adjusted R2 value of 0.945, an ACI score of 2634.683 and a GCV score of 11.738. Everything looks good, however, the visualization of the plot shown below indicates that the model is probably overfitting:

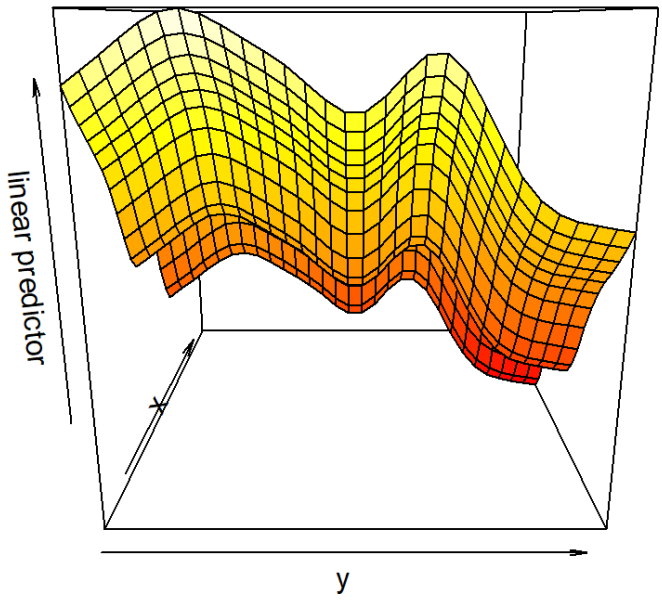


The response variable z seems to be too sensitive to the explanatory variables x and y. Therefore, although the model has a high R^2 value and a low ACI score, it is probably just a result of overfitting.

Generalized additive model (normal family)

```
seis.gam.normal <- gam(z ~ s(y) + s(x), data = seis)
```

Generalized additive model is used to model a response variable which is not from normal Gaussian distribution and that the response variable depends linearly on unknown functions. In this case, the normal family with identity link function is selected. The gam function from the mgcv package is again applied. The model yielded the same adjusted R^2, AIC and GCV as the bivariate spline.

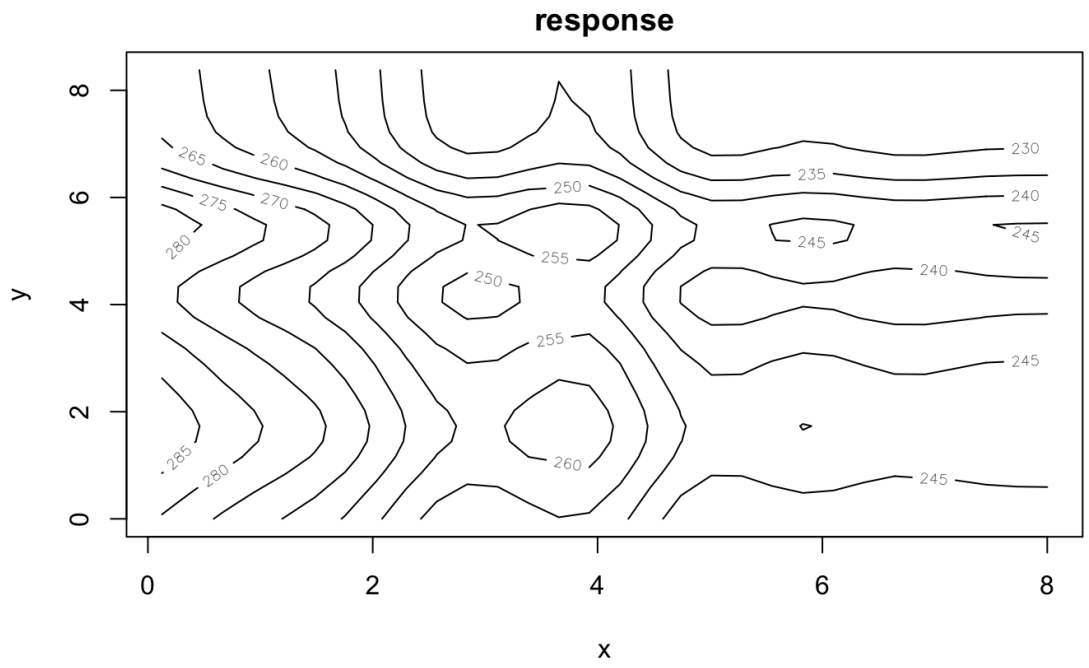


The visualized fitting is shown on the left side. It can be observed that although the fitting is smooth, it does not provide enough fitting of the raw data.

Generalized additive model (gamma family) - Recommended

```
seis.gam.gamma <- gam(z ~ s(y) + s(x), data = seis, family = Gamma(link = "log"))
```

Since the normal family do not provide better prediction as compared to other models, the GAM with gamma family is attempted. The model do not yield significant improvement in terms of R square and AIC. However, it obtained a GCV (generalized cross validation) of 0.00084 which is the lowest of among all other tested models. This means that with the similar performance in training, it provides long run prediction with much more confident error. The thin plate spline model performs well in terms of R2 and AIC. However, the high GCV implies that it might overfit which leads to high variance of the model thus not useful for long run prediction. Considering the variance-bias trade off, the GAM with gamma family is considered the best model. Below is the contour map based on that and is expected to provide guidance for geologists.



Summary of Model Outputs

Output	Multiple Linear Regression	Bivariate Spline	Thin-plate Regression	GAM with Normal Family	*GAM with Gamma Family
Adj. R^2	0.4695	0.676	0.947	0.676	0.678
AIC	3676.541	3442.686	2623.19	3442.686	3432.615
GCV	N.A.	54.062	11.575	54.062	0.00084565

Conclusion

GAM with Gamma Family is the final recommended model. It is chosen not only because it offers decent adjusted R^2 and AIC values among all the models, but more importantly because it tends to provide a more reliable long-term predictability according to its small GCV score.