FINAL PROJECT-Question 2

SNIGDHA PEDDI

INTRODUCTION

This data set consists of the log likelihood ratios for pairwise comparison of fourty inks obtained by paramertisation of spectra from microspectrometry in the visible range. Ten replicates of the fourty inks were parametrised and log likelihood ratios were calculated which is like a similarity score and is called Forensic likelihood ratio.The pairwise comparisons were calculated under propositions implying that any ink from the evidence document came from same pen vs it came from some other pen. The three univariate LR's corresponding to the "x","y","Z" measures of the color are part of this data set.The data set consists of 6 variables, Comparison of interest (1-820),log values of omnibus likelihood ratio(Dependent variable),Type of comparison("wi"-within source comparison and "bw"- between source comparison),Log likelihood ratio of x color variable,Log likelihood ratio of y color variable,Log likelihood ratio of z color variable.The relationship between these univariate LR's and omnibus LR constructed by Dr. Saunders will be analyzed.

ANALYSIS AND DISCUSSION

Exploratory Data Analysis The Comma Separated File is imported.The dimensions of the data set and presence of missing values were verified.And the summary of the data set give a basic idea of values in the data set (mean ,median values etc.).

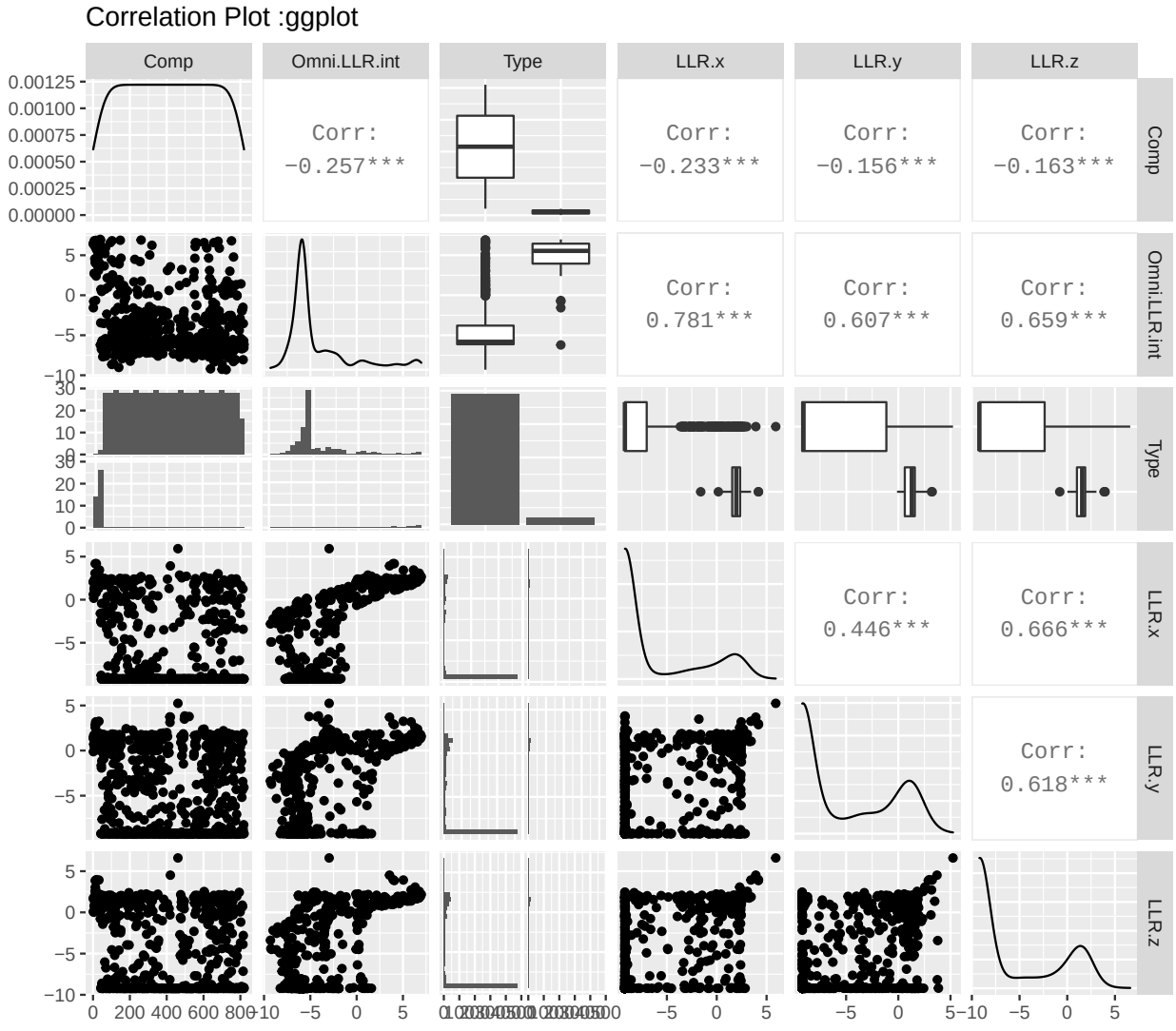
```
## [1] 820   6
```

```
##
##  Number of missing values in dataset: 0
```

```
##       Comp        Omni.LLR.int    Type         LLR.x            LLR.y
##  Min.   :  1.0   Min.   :-9.309   bw:780   Min.   :-9.210   Min.   :-9.2103
##  1st Qu.:205.8   1st Qu.:-6.109   wi: 40   1st Qu.:-9.210   1st Qu.:-9.2103
##  Median :410.5   Median :-5.822            Median :-9.210   Median :-8.8836
##  Mean   :410.5   Mean   :-4.130            Mean   :-6.529   Mean   :-5.3592
##  3rd Qu.:615.2   3rd Qu.:-3.253            3rd Qu.:-3.431   3rd Qu.:-0.1435
##  Max.   :820.0   Max.   : 6.982            Max.   : 5.922   Max.   : 5.2407
##      LLR.z
##  Min.   :-9.2103
##  1st Qu.:-9.2103
##  Median :-9.2092
##  Mean   :-5.8425
##  3rd Qu.:-0.5485
##  Max.   : 6.6033
```

Correlation between the variables is reviewed by plotting the data using ggpairs function.From the plot it is clear that the three LR's do not have a linear relation to omnibus LR. To study the relationship between these variables inear models with spline smoothing will be fit and analyzed.Generalized Additive Models(GAM)

will be fit for this purpose. GAM's give us a framework to model flexible nonlinear relationships and use basis functions to make smooth curves.



Fitting the Models and Discussion

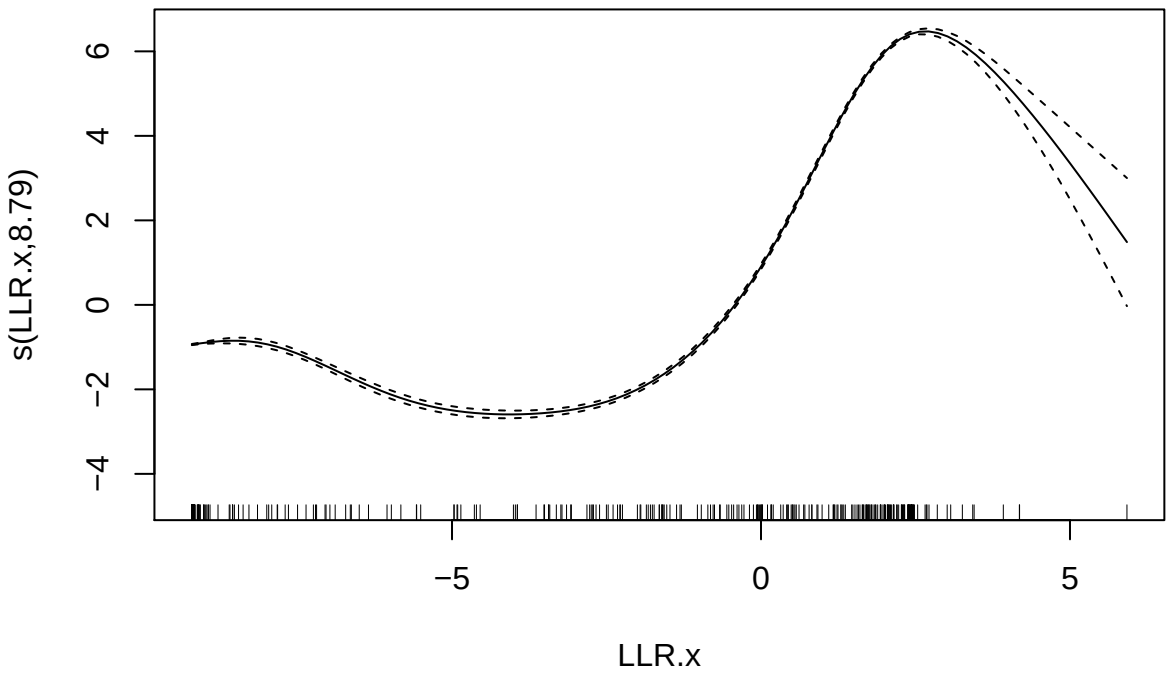
To check if there is any relationship between variables, aov() function is used. The lower pvalues indicate that at 95% confidence interval all the variables are related significantly.

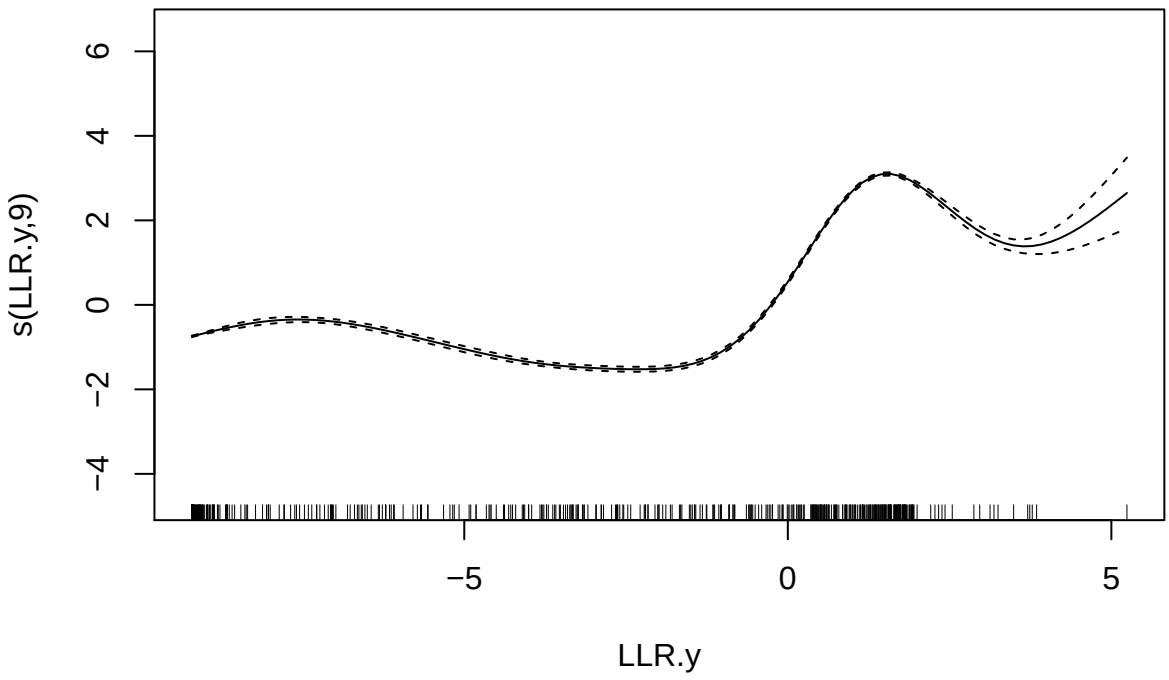
```
## Model:

## Omni.LLR.int ~ (LLR.x) + (LLR.y) + (LLR.z) + Type

##              Df Sum Sq Mean Sq  F value  Pr(>F)
## LLR.x         1   6516    6516 1833.509 < 2e-16 ***
## LLR.y         1    889     889  250.080 < 2e-16 ***
## LLR.z         1     31      31    8.644 0.00337 **
## Type          1    341     341   96.082 < 2e-16 ***
## Residuals   815   2896       4
```

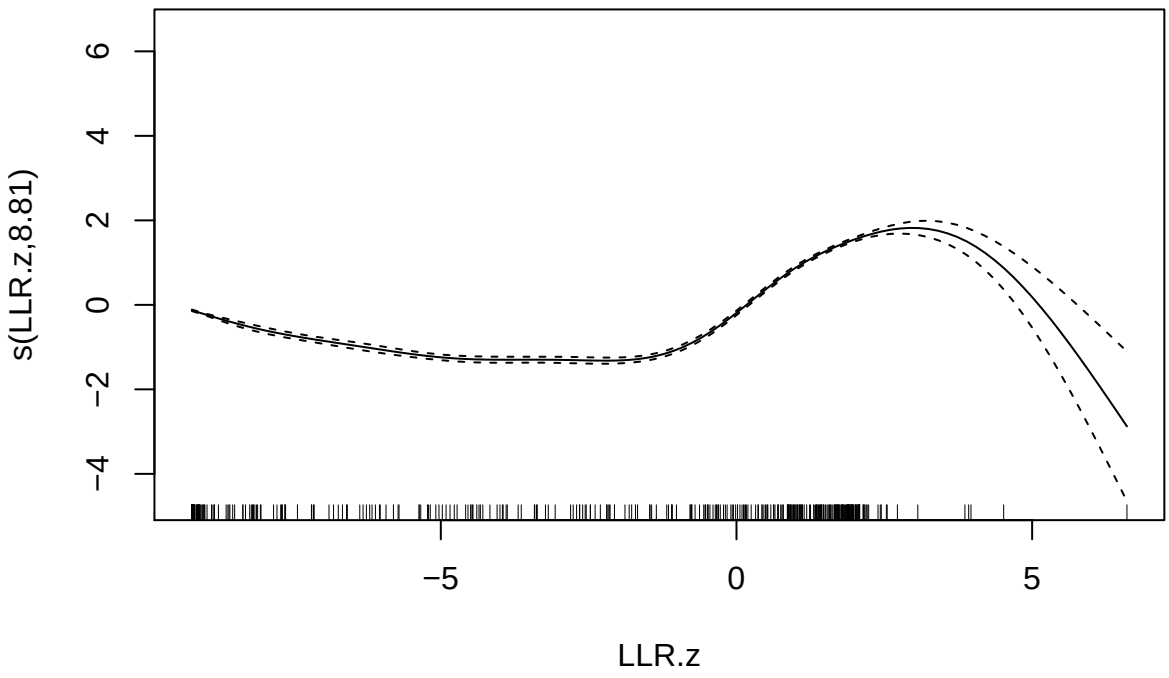
```
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Linear regression models were fit with splines to explain the relationship between Omnibus LR and other LR's. A GAM model is fit between the 3 univariate LR's and omnibus LR without smoothing. The p values indicate a strong relationship between the variables. However the analysis is contuned to verify the effect of spline on the variables. A GAM model is fit between the 3 univariate LR's and omnibus LR. The p values indicate a strong relationship between the variables. From the plots it is clear that they have a non-linear relationship. The omnibus LR values increases slowly with increase in the LLR.x and around a value of 3 then tend to decrease.The omni bus values increase slowly with increase in LLR.y LR around -0.5 and then tend to reduce around value of 2 and then slowly increase toward the higher likelihood values.Similar non linear patter is observed with LLR.z where values slowly decrease and start increasing around -0.5 and decrease around likelihood values of 3.

```
## Model:

## Omni.LLR.int ~ LLR.x + LLR.y + LLR.z

## Table of Linear terms and Intercept:

##               Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) 0.69733977 0.13143764  5.305480 1.448761e-07
## LLR.x       0.49272976 0.02134419 23.084963 3.824936e-91
## LLR.y       0.23012067 0.01951992 11.789016 9.844000e-30
## LLR.z       0.06448805 0.02317641  2.782487 5.518971e-03

## Model:

## Omni.LLR.int ~ s(LLR.x) + s(LLR.y) + s(LLR.z)

## Table of Linear terms and Intercept:

##              Estimate  Std. Error   t value Pr(>|t|)
## (Intercept) -4.129599 0.006572502 -628.3146        0

##
##
##  Table of Smooth Terms :

##                edf   Ref.df         F p-value
## s(LLR.x) 8.792598 8.975549 10272.519       0
## s(LLR.y) 9.000000 9.000000  4058.185       0
## s(LLR.z) 8.807416 8.977645  1064.881       0
```


s(LLR.x,8.79)
6
4
2
0
−2
−4
−5
0
5
LLR.x

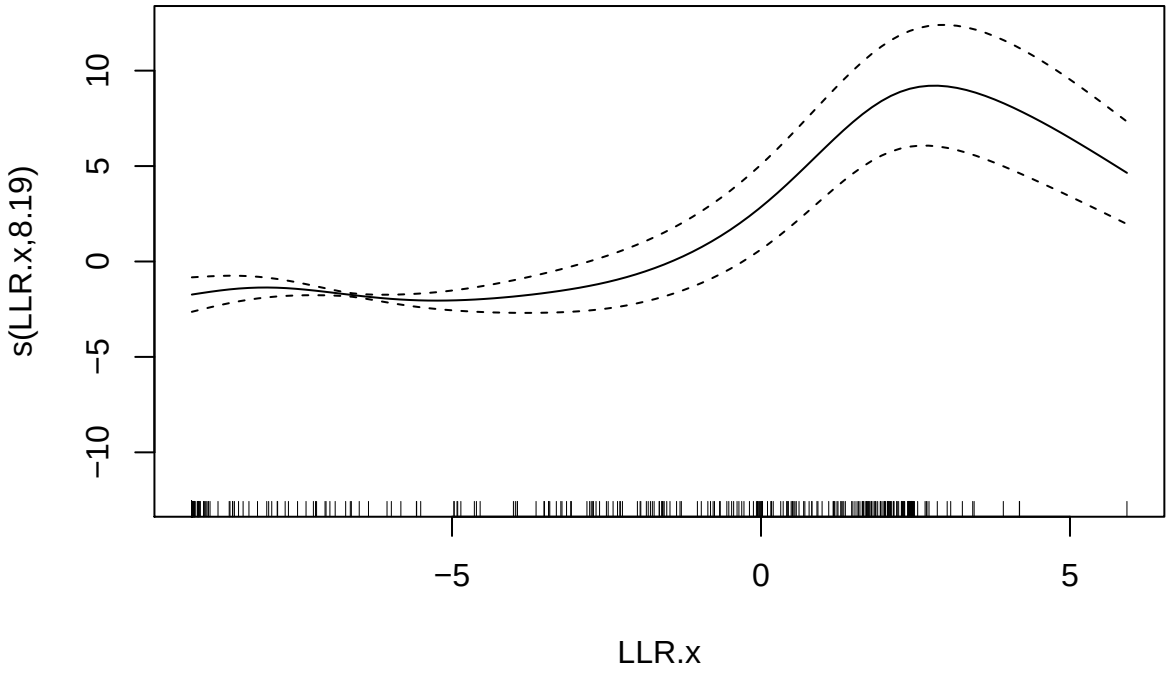

s(LLR.y,9)
6
4
2
0
−2
−4
−5
0
5
LLR.y

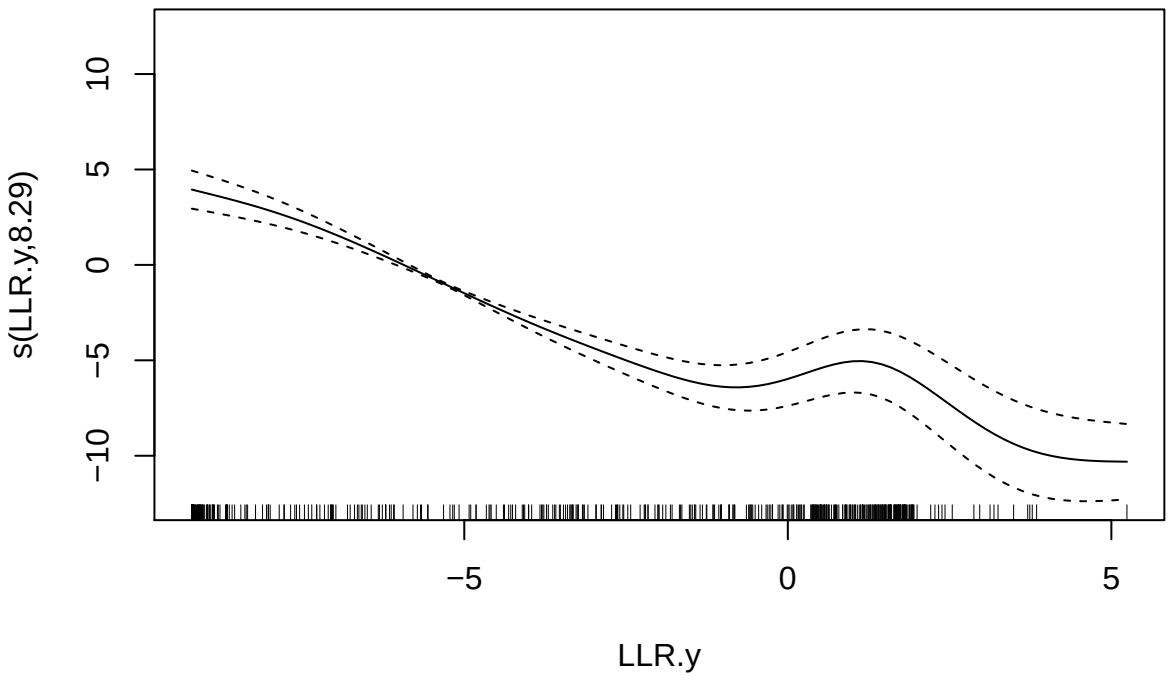


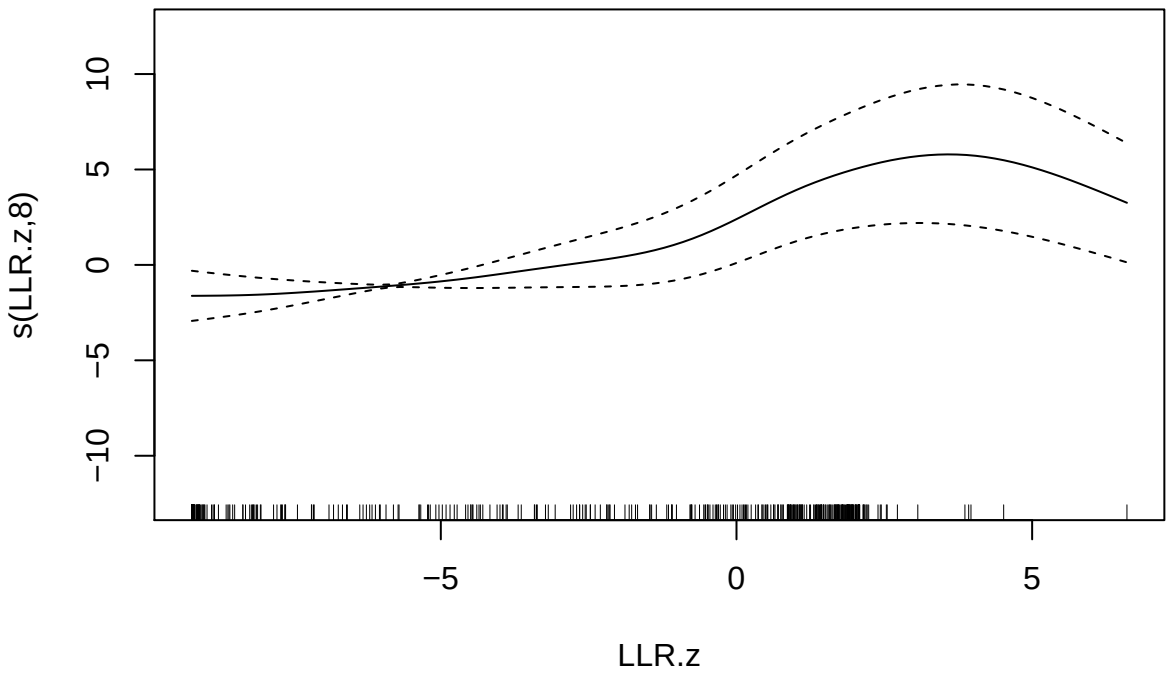
Added an interaction term of Type variable with the LLR's show that variables have significant relationship with the type of comparison if they are within source comparison or between source comparison.In presence of the interaction term, though the relationship is non-linear the values are spread out and different for LLR.y where the likelihood values decrease until -0.5 significantly and then tend to increase and then decrease around value of 2.

```
## Model:

## Omni.LLR.int ~ s(LLR.x) + s(LLR.y) + s(LLR.z) + Type:(LLR.x +
##     LLR.y + LLR.z)

## Table of Linear terms and Intercept:

##                Estimate Std. Error    t value     Pr(>|t|)
## (Intercept)  -2.1385276  0.1031090 -20.740452 1.299262e-76
## Typebw:LLR.x -0.2973532  0.1692159  -1.757242 7.926417e-02
## Typewi:LLR.x -0.3067158  0.1862945  -1.646403 1.000788e-01
## Typebw:LLR.y  1.2174965  0.1298237   9.378076 6.931496e-20
## Typewi:LLR.y  1.2413634  0.1718933   7.221711 1.209318e-12
## Typebw:LLR.z -0.4430384  0.1956750  -2.264154 2.383514e-02
## Typewi:LLR.z -0.3997795  0.2711689  -1.474282 1.408043e-01

##
##
##  Table of Smooth Terms :
```

```
##                edf   Ref.df         F p-value
## s(LLR.x) 8.188192 8.378370 4878.3471       0
## s(LLR.y) 8.289337 8.289337 2246.6360       0
## s(LLR.z) 7.997492 8.264116  917.1259       0
```




s(LLR.y,8.29)
10
5
0
−5
−10
−5
0
5
LLR.y



GAM model with LR's from x,y,z measures with their spline interaction term indicate that all the terms are significant at 95% confidence interval.

```
## Model:

## Omni.LLR.int ~ LLR.x + LLR.y + LLR.z + s(LLR.x + LLR.y + LLR.z)

## Table of Linear terms and Intercept:

##                Estimate Std. Error   t value     Pr(>|t|)
## (Intercept) -5.13567426 0.77994234 -6.584685 8.196769e-11
## LLR.x       -0.42623173 0.12125935 -3.515042 4.641200e-04
## LLR.y        0.22274370 0.01193570 18.661966 6.483966e-65
## LLR.z        0.09977597 0.01436669  6.944953 7.778854e-12

##
##
##  Table of Smooth Terms :

##               edf   Ref.df        F       p-value
## s(LLR.x) 7.573358 8.364629 170.9272 9.804294e-259
```

To check if the addition of the interaction term has improved the model performance anova test is performed between models. A p value of 0.2527 indicate that addition of interaction term did not the model performance. Hence, GAM model with no interaction term will be used for further analysis.The GCV score or the

generalized cross validation score is an estimate of mean square error of the LOOCV process.Comparing the GCV values and the Adjusted R square values also suggest that the model with the LR's of x,y,z measures with smoothing is best out of the four models.

```
## Analysis of Deviance Table
##
## Model 1: Omni.LLR.int ~ s(LLR.x) + s(LLR.y) + s(LLR.z)
## Model 2: Omni.LLR.int ~ s(LLR.x) + s(LLR.y) + s(LLR.z) + Type:(LLR.x +
##     LLR.y + LLR.z)
##   Resid. Df Resid. Dev    Df Deviance Pr(>Chi)
## 1    792.05     28.069
## 2    789.09     27.926 2.957  0.14242   0.2527
```

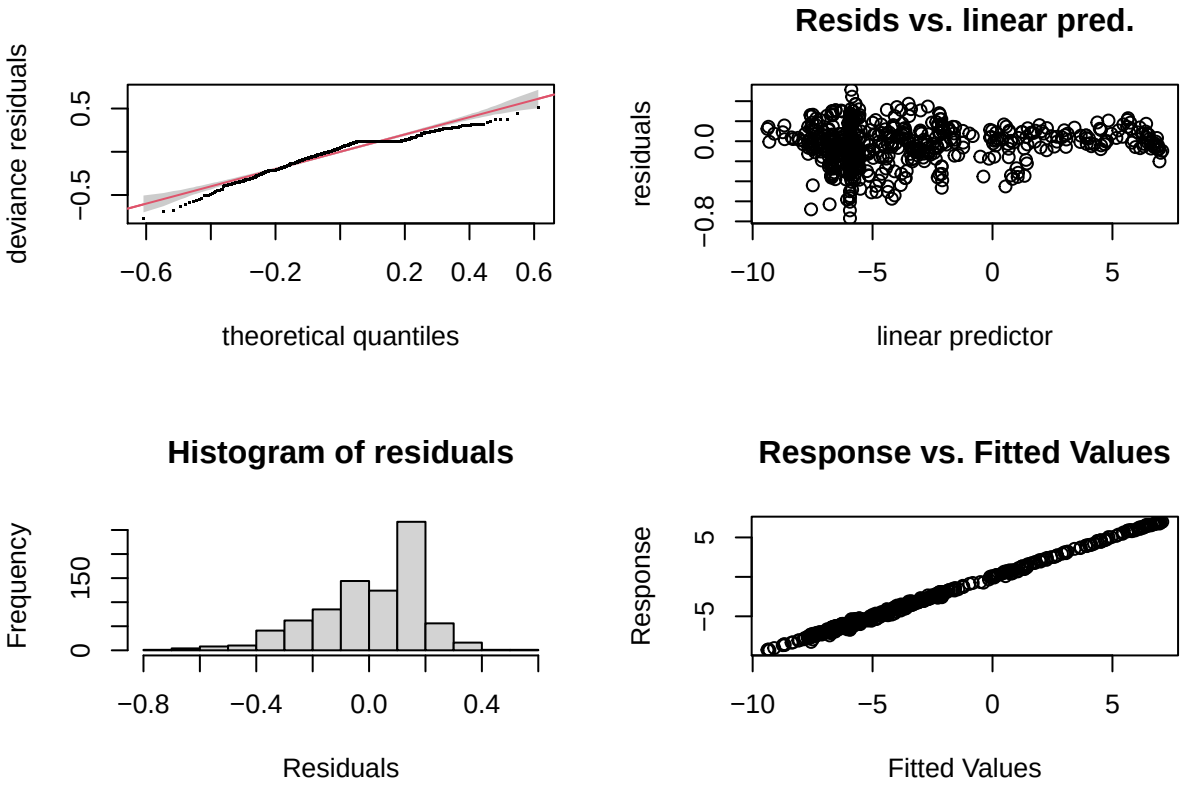
```
## Comaprison of GCV of models:
```

```
##        GCV_w.o.Spline GCV_w.Smoothing GCV_w.interaction.terms
## GCV.Cp       3.987425      0.03665596              0.03673405
##        GCV_interaction.smooting
## GCV.Cp                 1.495063
```

```
## Comaprison of Adjusted R-square of models:
```

```
##   R2_w.o.Spline R2_w.Smoothing R2_w.interaction.terms R2_interaction.smooting
## 1     0.6955261       0.997282               0.997286               0.8868015
```

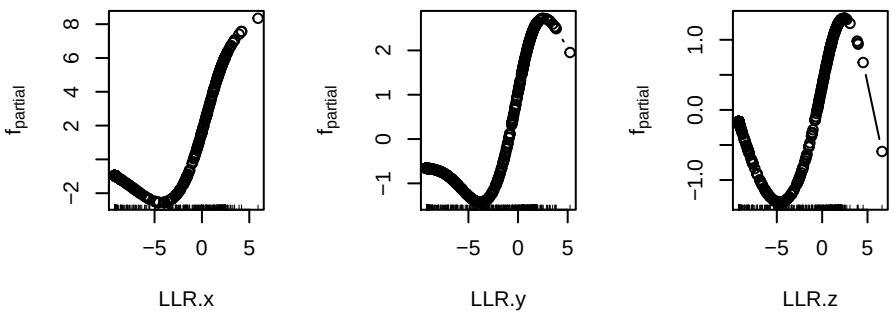
The Quantile plot show that most of the points do not fall within the confidence interval which is a concern.The residual vs.linear predictor plot shows that the spread widening out which is of concern.The histogram looks guassian and left skewed.The Response vs.Fitted values plot have linear fit indicating the model is performing good.The model has to be improved.




```
##
## Method: GCV   Optimizer: magic
## Smoothing parameter selection converged after 19 iterations.
## The RMS GCV score gradient at convergence was 9.986767e-08 .
## The Hessian was positive definite.
## Model rank =  28 / 28
##
## Basis dimension (k) checking results. Low p-value (k-index<1) may
## indicate that k is too low, especially if edf is close to k'.
##
##            k'  edf k-index p-value
## s(LLR.x) 9.00 8.79    0.95    0.09 .
## s(LLR.y) 9.00 9.00    0.63  <2e-16 ***
## s(LLR.z) 9.00 8.81    0.48  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

GAMboost() function is used to fit the GAM model by likelihood based boosting.It provides the smooth function estimates of covariates along with confidence bands and DF.The AIC suggests that the boosting algorithm should be stopped after 100 iterations.The Type and comparison of interest variables are removed by the AIC method and the resulting AIC is very low indicating this model is performing better.The plots of the models shows that the variables are smoothed enough to have a linear relationship with the omnibus LR. The number of degrees of freedom are 10 and are more than the GAM model indicating increased number of base learners and hence improved performance.All the plots show a linear relationship from around -4 likelihood values and until around 3.And the Omnibus LR values decrease with increase in the LR values of other 3 measures towards lower and upper end.

```
## [1] 0.1764985
## Optimal number of boosting iterations: 100
## Degrees of freedom (for mstop = 100): 9.986192

##
##   Model-based Boosting
##
## Call:
## gamboost(formula = Omni.LLR.int ~ ., data = omni)
##
##
##   Squared Error (Regression)
##
## Loss function: (y - f)^2
##
##
## Number of boosting iterations: mstop = 100
## Step size:  0.1
## Offset:  -4.129599
## Number of baselearners:  5
##
## Selection frequencies:
## bbs(LLR.y, df = dfbase) bbs(LLR.x, df = dfbase) bbs(LLR.z, df = dfbase)
##                    0.40                    0.35                    0.25

##
##
##  AIC of model: 0.1764985
```



To confirm the improved model performance analysis of variance is performed between GAM model and GAMboost model.The lower P values indicate that the GAMboost model is better than the GAM model at 95% confidence interval.

```
##
## Family: gaussian
## Link function: identity
##
## Formula:
## Omni.LLR.int ~ s(LLR.x) + s(LLR.y) + s(LLR.z)
##
## Approximate significance of smooth terms:
##            edf Ref.df     F p-value
## s(LLR.x) 8.793  8.976 10273  <2e-16
## s(LLR.y) 9.000  9.000  4058  <2e-16
## s(LLR.z) 8.807  8.978  1065  <2e-16
```

CONCLUSION

Relationship between the Omnibus LR and the three LR's of x,y,z measures is studied.Initial investigation shows a non-linear relationship between the variables and omnibus LR.Generalized Additive Models were fit with smoothing variables of these measures. An Adjusted R square of 0.997 and GCV of 0.036 indicate that the GAM model is performing well. However, the Quantile plot of the model indicate that most of the values are not within the confidence interval and there is room for model improvement. Generalized Additive Model by likelihood based boosting(GAMboost) is fit. The model indicates an increased number of degrees

of freedom which in turn indicates increased number of parameters involved and hence improved model.To confirm that the GAmboost model has improved performance over the GAM model anova of the models is performed and the lower p values show that the GAMboost model has improved performance.The plot of the model shows a linear relationship between the x,y,z measures to the omnibus LR. Between likelihood values of -0.4 and around 3 with increase in the values of the measures the omnibus LR increases whereas towards the lower and higher side, omnibus LR decreases with increase in other three measures.

REFERENCES

- Snigdha Peddi,*Stat 601 Homework Assignment 6*
- Snigdha Peddi,*Stat 601 Homework Assignment 9*
- Lecture by Gavin Simpson,*Introduction to Generalized Additive Models with R and mgcv*,July 30,2020,(https://www.youtube.com/watch?v=sgw4cu8hrZM)
- Lecture from DataCamp,*Introduction to Generalized Additive Models*,April 13,2020,(https://www.youtube.com/watch?v=6V_VvweZkoI)