Final Project 2

Mohamed Ahmed

11/26/2020

Introduction:

The purpose of this report is to analyze and explain the relationship between four forensic likelihood ratios for different ink samples. We are given three univariate likelihood ratios linking to the "X"," Y", and" Z" Which are measures of color which were created. The second component of the analysis is to an omnibus LR that (Dr. Saunders) created. We want to analyze the given data and figure out if there is a relationship between the marginal LR's (LLR.x, LLR.y, and LLR.z) and the Omnibus LR, and explain the relationship if there any.

Exaplantory Data Analysis:

Before we started out analysis, the first thing we did was perform some initial data inspection. We decided to drop the column comparison. We constructed a matrix of plots to visualize the data. Then we used constructed histograms to look at the distribution of the data. **Figure 2** shows that the covariates are very skewed to the right. We can see that the variable of interest omnibus LLR is skewed to the right as well, and it is suspected that is normally distributed. To understand more about the how covariates are correlated, we constructed a matrix **(Table 2)** that has the correlation coefficients. we can see that the correlation between LLR.x and LLR.y is **0.446** which is a bit weaker positive correlation compared to the positive correlation between LLR.x and LLR.z **0.666** which is a moderate relationship. also, the relationship between LLR.y and LLR.z is a positively moderate relationship with a correlation coefficient of **0.618**. Form **figure 1**, there is a very clear curved upward positive trend between omnibus and LLR.x. It could be a linear relationship that requires a polynomial term. The relationship between omnibus and the other two covariates seem to be constant for a bit and then there is an upward trend towards the end. By looking at **figure 1**, the scatter plots for Omnibus Likelihood ratio and each of the Log Likelihood ratio of each color(X,Y,Z) shows that there is a potential outlier in middle right part of the graphs which belong to type *bw*. We are not sure if a linear relationship exists between the variable of interest and the covariates LLR.x; therefore, we will start our analysis by fitting a simple linear regression model.

Model

Linear Regression Models:

initially, we assumed a linear relationship, then We started our analysis by constructing multiple simple linear regression models with one explanatory variable for each model to check if there is a linear relationship between the variable of interest, in this case, LLR.omnibus and the covariates. in **figure 3** , We have fitted the simple linear regression line in each plot, to check if a model captures the relationship between the response variable and each of the covariates but we can clearly see that the linear models do not fully capture the trend and do not explain the relationship between The explanatory variables and omnibus. The only thing that was noticeable was the fact that that the relationship between Omnibus LLR and LLR.x looked like it could be explained by a polynomial model.

Polynomial Model:

Since we could not see any linear relationship between the response variable and the covariates LLR(x,y,z), We tried a more general Sami-parametric lowess smoother for each covariate to try to capture the relationship between the response variable and the covariates. we can see from **Figure 4** that the relationship is non-linear between the response variable omnibus.LLR and the covariates. since we have more than one explanatory variable, using generalized additive is more adequate than using a single scatter plot smoother.

Generlized Additive Model:

Since we have a non-linear relationship, we decided to use a generalized additive model because it can help us model the non-linear relationship between the response variable and the covariates using smoothers.

GAM Variable Selection:

Before we try to select a model, we assumed normal errors. We used a simple logical way to select the variables that will enter the model as parametric and non-parametric variables. initially, A full model was built that included all covariates as smoothed terms except for Type which was entered into the model as a parametric term. we knew that the three covariates LLR.x, LLR.y, and LLR.z need smoothing from our previous models. Next step was to remove a variable at time and see how Generalized cross validation value changes. GCV is considered the most logically consistent method to use for variable selection process. After constructing many models and comparing them using GCv, the model with the lowest GCV **0.0366034(Table 3)** was the full model. However, we decided to select **Model 2** as the selected model because the variable **Type** was dropped because its P-value was below the significance level $\alpha = 0.5$. The GCV **0.0366560** for the selected model and the optimal model was insignificant. Table 2 displays all the models that were constructed along with more information about each model.

GAM Model Results:

The parametric term is the intercept, and its p-value is significant at an $\alpha = 0.5$. **Table 4** shows the results for the parametric terms. **Table 5** shows the smooth components of our model regarding the marginal LR's (i.e. LLR.x, LLR.y, and LLR.z) and its relationship with the Omnibus LR which suggests that all the smoothed marginal LR"s are significant at an $\alpha = 0.5$. The GCV score is an estimate of mean square prediction error which can be found in **Table 6**. For our model, our **GCV = 0.036656** which is the second lowest out of all the models we have attempted to build. Also, our model's **AIC = -382.9581** is the second lowest out of all the models we built. we looked at the partial contributions of the covariates. Each plot shows the effects covariates have on the response. From **Figure 5a**, the first plot shows that LLR.x stays flat and constant until about zero and then it spikes, and finally decreases after the value of **3**. **Figure 5b** shows that LLR.y does not have much of an effect until**zero** then it increases for a bit but then decreases quickly. From **Figure 5c**,we see that LLR.z does not have an effect up until zero then it increases and finally, it decreases at the end. It seems that LLR.x has the most effect of the response variable.

GAM Model Diagnostics:

The residuals plot show that error distributed randomly around zero. There could be a pattern in the residuals which is not a good sign. Also, the histogram shows a somehow a normal distribution and the response vs fitted plot shows a linear relationship.

Conclusion:

We have analyzed the LLR data to check if there is a relationship between the marginal LLRs (x, y, z) and omnibus LLR. We have explored the data to gain a better understanding then we tried to use different parametric and non-parametric methods to find if there is a relationship and if there is one figure out the nature of the relationship. we can conclude that there is a relationship between the marginal LLRs (x, y, z) and omnibus likelihood ratio and that the nature of the relationship is non-linear.

Figures and Tables

Data

Symbol	Description
$Comp$	Comparison of interest (from 1 to 820).
$Omni.LLR.int$	numeric values of for the Log of the Omnibus likelihood ratio, this is the response variable we are interested in.
$Type$	Type of comparison, either "wi" for within-source comparison or "bw" for between-source comparison
$LLR.x$	The Log Likelihood ratio for the X color variable

$LLR.y$	The Log Likelihood ratio for the Y color variable
$LLR.z$	The Log Likelihood ratio for the Z color variable

```
## Data Dimensions

## [1] 820   5
```

Table 1: Statistical Summary

	Vname	Group	nNeg	nZero	nPos	NA_Value	Per_of_Missing	min	max	mean	median	IQR	nOutliers
2	LLR.x	All	680	0	140	0	0	-9.21	5.92	-6.53	-9.21	5.78	1
3	LLR.y	All	620	0	200	0	0	-9.21	5.24	-5.36	-8.88	9.07	0
4	LLR.z	All	635	0	185	0	0	-9.21	6.60	-5.84	-9.21	8.66	0
1	Omni.LLR.int	All	700	0	120	0	0	-9.31	6.98	-4.13	-5.82	2.86	93

Figure 1: Matrix Of Plots

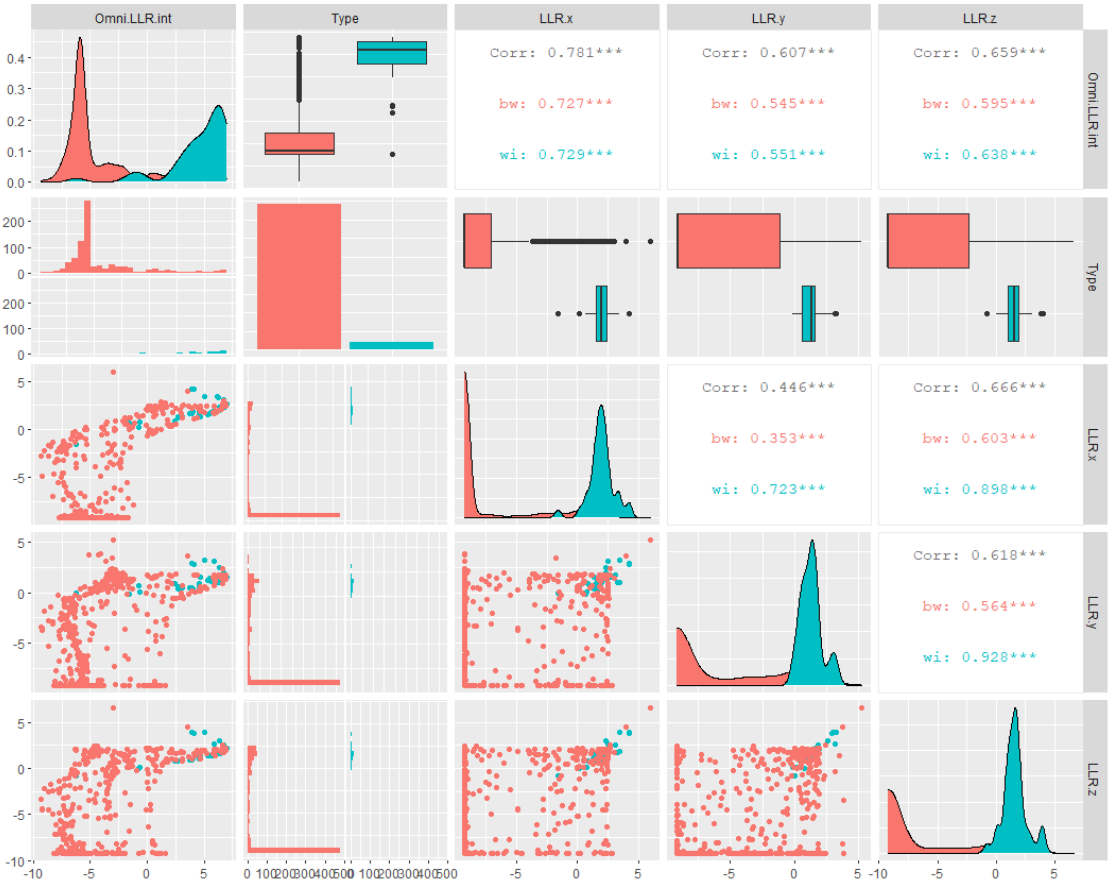


Figure 2: Histograms for continous variables

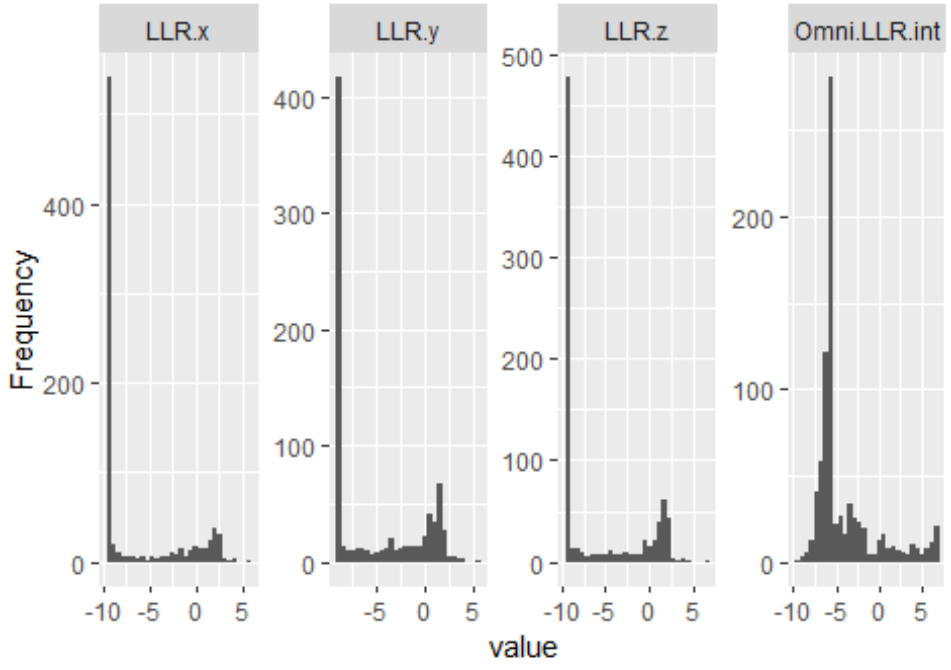


Table 2: Variables' Correlation Table

	Omni.LLR.int	LLR.x	LLR.y	LLR.z
Omni.LLR.int	1.000	0.781	0.607	0.659
LLR.x	0.781	1.000	0.446	0.666
LLR.y	0.607	0.446	1.000	0.618
LLR.z	0.659	0.666	0.618	1.000

Figure 3a: Fitting a Simple Linear Model Line

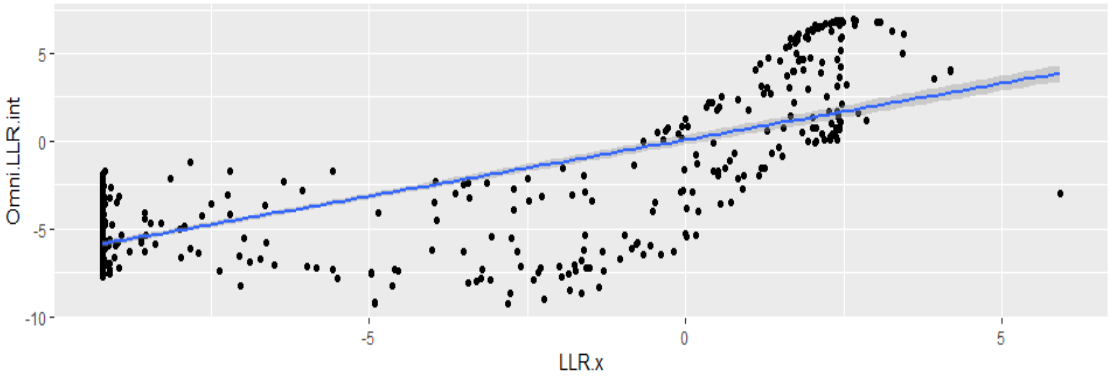


Figure 3b: Fitting a Simple Linear Model Line

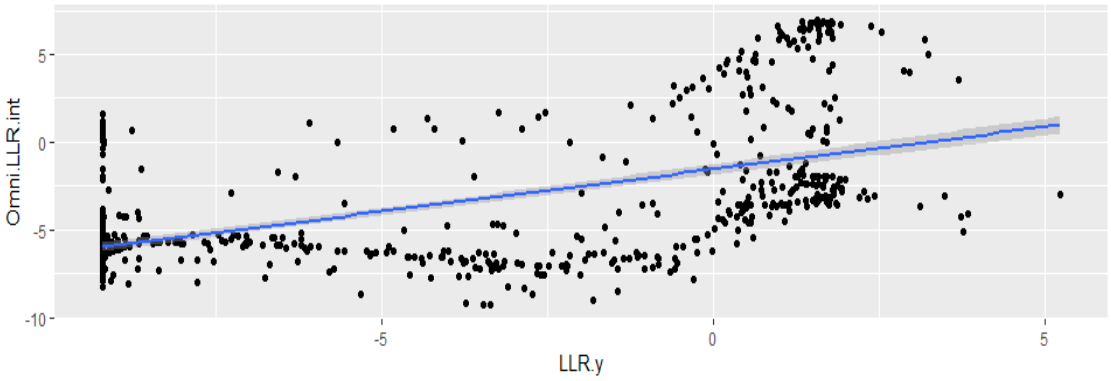


Figure 3c: Fitting a Simple Linear Model Line

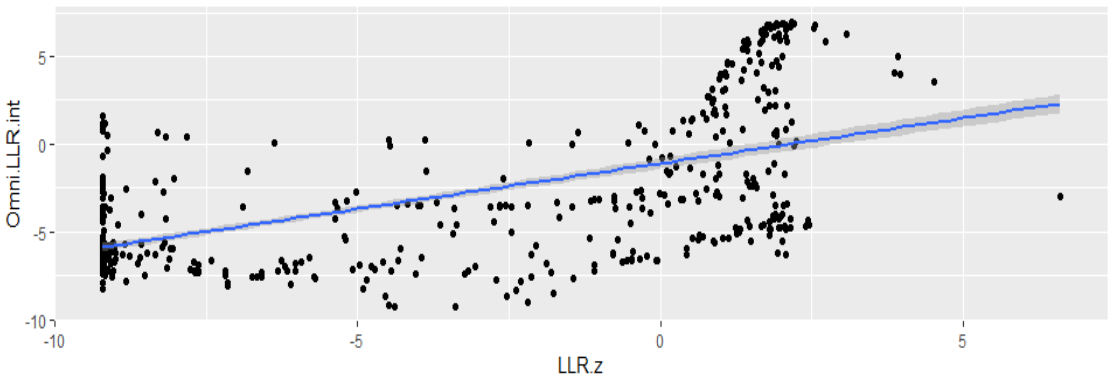


Figure 4: Lowess smoother for each Explantory variable

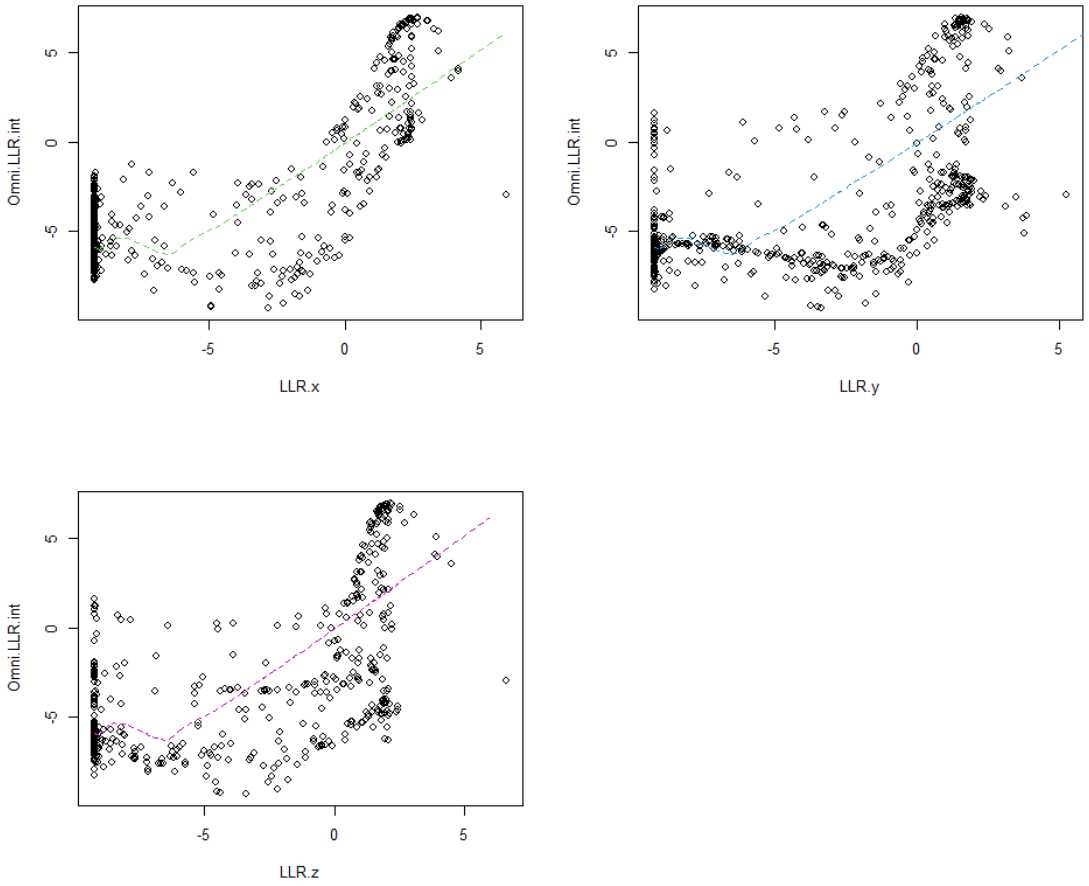


Table 3: Models' selection Information

Model	GCV	AIC	TotalModelDef
Full-Model	0.0366034	-382.9581	27.60001
Model-2	0.0366560	-384.2012	28.54758
Model-3	3.7470960	3412.0128	14.26142
Model-4	1.6018690	2715.1343	15.19157
Model-5	0.4605303	1692.7887	19.30330
Model-6	0.4686593	1707.1945	18.06096
Model-7	1.6757570	2752.1475	14.20656
Model-8	4.2225130	3509.9376	14.91402

Table 4: Parametric Terms significance of the GAM Model

	Estimate	Std. Error	t value	Pr(>|t|)
(Intercept)	-4.129599	0.0065725	-628.3146	0

Table 5: Smooth terms Significance

	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

Table 6: Information on the Selected Model

	GCV	AIC	Total.Model.DoF
GCV.Cp	0.036656	-382.9581	27.60001

Figure 5a: Partial Contribution of LLR.x

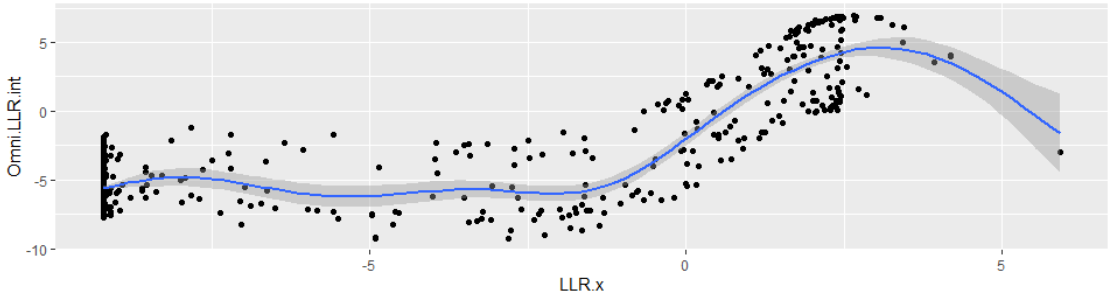


Figure 5b: Partial Contribution of LLR.y

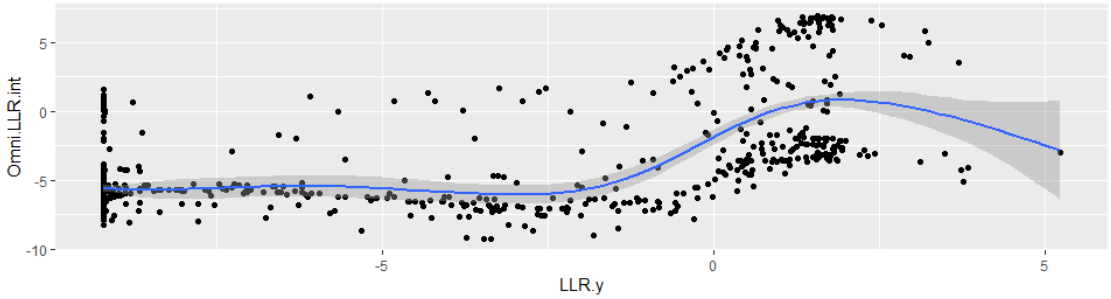
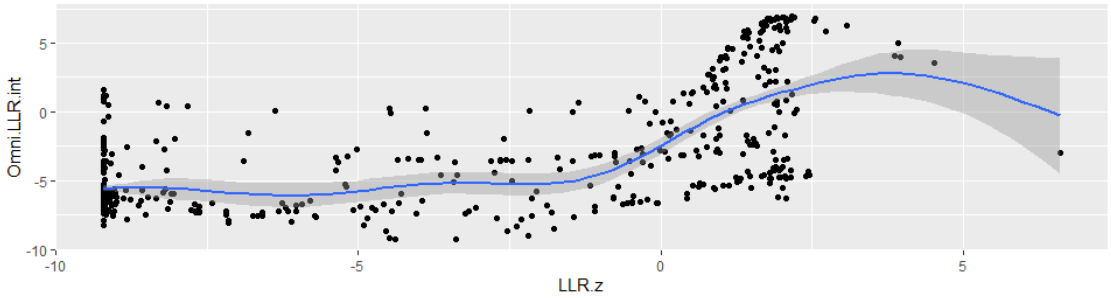
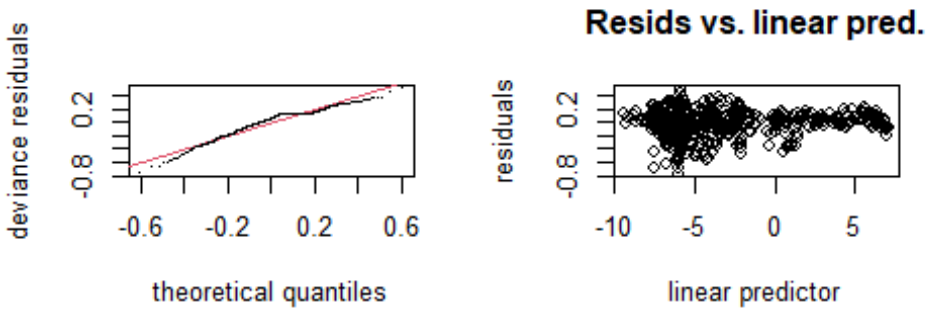
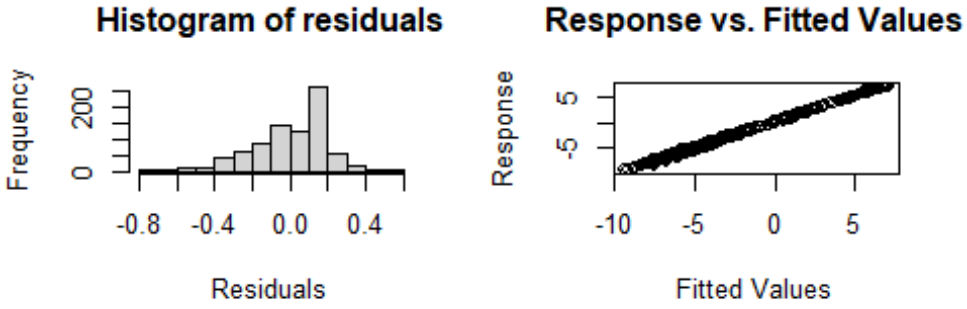


Figure 50c: Partial Contribution of LLR.z








```
##
## 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   0.065 .
## 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
```

Citations

Generalized Additive Models /. (2019, February 17). Retrieved December 07, 2020, from https://m-clark.github.io/generalized-additive-models/application.html

Everitt, Brian, and Torsten Hothorn. A Handbook of Statistical Analyses Using n SECOND EDITION. Taylor and Francis Group LLC, 2010.