Determining relationship between the marginal LR and the omnibus LR

Yamuna Dhungana

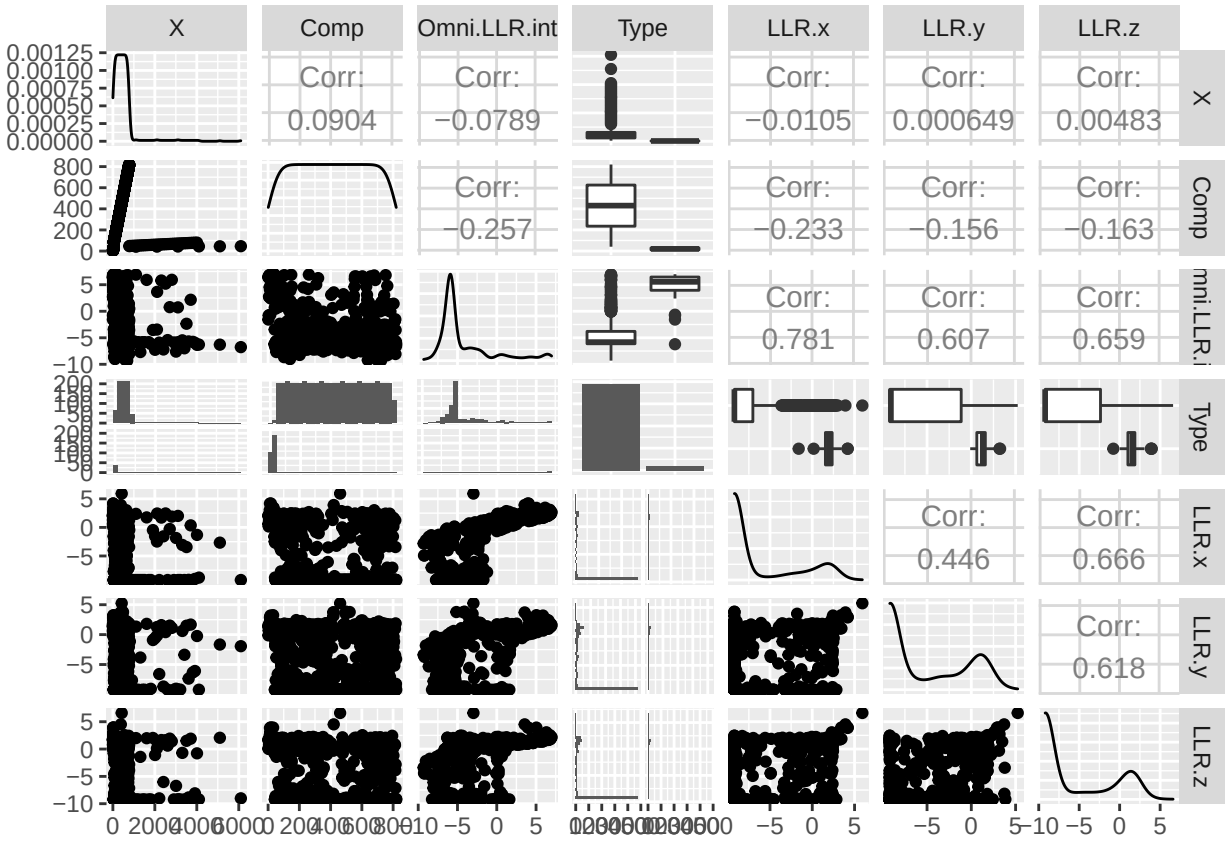
The data is taken from "The evidential value of micro spectrophotometry measurements made for pen inks", by Martyna et al. (2013). The author was focused on developing a type of similarity score, between two ink samples, known as a forensic likelihood ratio. The sample for the data was taken from the 40 inks. 36 of the 40 ins was from the ballpoint pen, and 4 of the ink is from the gel pen. They came either from the Polish market or were gifted presented to the institute of Forensic Research. Lines were created by drawing on the white printing paper. The fragment of paper was viewed on the microscope. The colors used for the analysis were red, green, and blue. The data are given in the dat.LLR.int file has a total of 820 samples with the six variables. The variables described as: 1. Comp : comparison of interest (from 1 to 820) 2. Omni.LLR.int : numeric values of for the Log of the Omnibus likelihood ratio. 3. Type : Type of comparison, either "wi" for within-source comparison or "bw" for between-source comparison 4. LLR.x : The Log Likelihood ratio for the X color variable 5. LLR.y : The Log Likelihood ratio for the Y color variable 6. LLR.z : The Log Likelihood ratio for the Z color variable

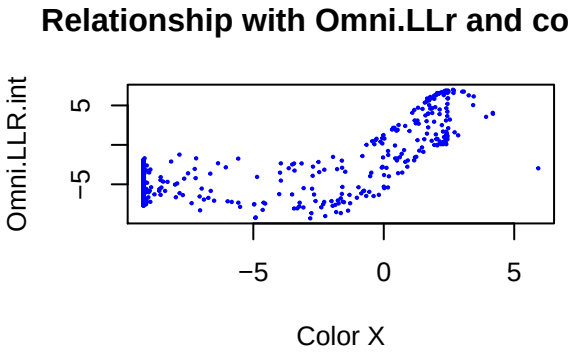
The point of interest in our project is to find the relationship between the marginal LR's (i.e. LLR.x, LLR.y, and LLR.z) and the Omnibus LR (i.e. Omni.LLR.int).

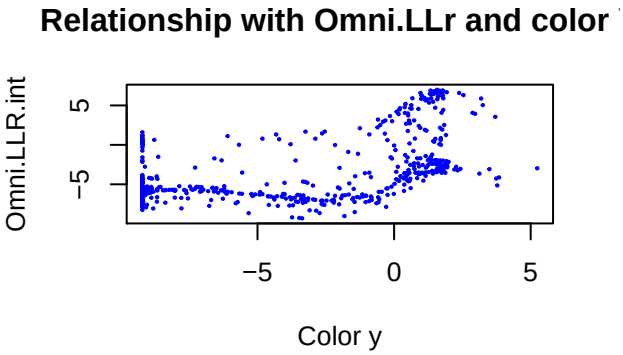
After loading the data, the values of the data are visualized. From the graphs, it looks like all the variables individually have some relationship with omnibus LR. All the variable seems to have a linear relationship(quadratic); however, we cannot say and assume to have it until we have some evidence. To start with, I have plotted a pair-wise plot and see their correlation plot with coefficients. For the correlation plot, the correlation between LLR.x, LLR.y, and LLR.z look moderately correlated with the 0.78, 0.67, and 0.659, respectively.

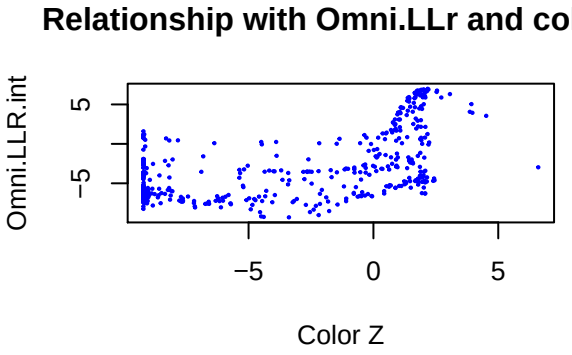
```
## Registered S3 method overwritten by 'GGally':
##   method from
##   +.gg   ggplot2
```

```
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
## `stat_bin()` using `bins = 30`. Pick better value with `binwidth`.
```


X
Comp
Omni.LLR.int
Type
LLR.x
LLR.y
LLR.z
Corr: 0.0904
Corr: −0.0789
Corr: −0.0105
Corr: 0.000649
Corr: 0.00483
Corr: −0.257
Corr: −0.233
Corr: −0.156
Corr: −0.163
Corr: 0.781
Corr: 0.607
Corr: 0.659
Corr: 0.446
Corr: 0.666
Corr: 0.618







Further, I have made a combination of different variables and different interactions. The model that follows in the project is a single variable, two variables (also variables with interactions), and three variables. I have plotted a total of 11 models for linear regression.

1. Linear regression

Firstly, I Fit the linear regression and viewed the p- values and R- squared. Then the MSE is calculated. Additionally, to make sure, the same process was repeated to see P-values and MSE for the quadratic models. Finally, I performed the Anova to test to see the better fit.

The model mentioned for a single variable is fitted with the LLR.x, LLR.y, and LLR.z with Omnibus.LLR individually as `Omni.LLR.int ~ LLR.x`, `Omni.LLR.int ~ LLR.y`, `Omni.LLR.int ~ LLR.z` In the two pair models, I combined these variables and also saw the interaction with these variables. The same method follows with the three variables. Although the table is made individually according to no of the variable used, these variables are combinedly compared. Mean square error is the average squared difference between the estimated values and the actual value. The low MSE is always better. Therefore, by looking at the table for the MSE of a model the decision is taken. Hence by the MSE, all the variables combined have the linear regression.

1.1 Linear regression for single variable

Table 1: MSE for linear regression model with single variable

Model-1	MOdel-2	model-3
5.07	5.07	7.366

1.2 Linear Regression for two pair

Table 2: MSE for linear regression model with two variables

Model-4	MOdel-5	model-6
3.986	6.527	4.621

Table 3: MSE for linear regression(interacting) model with two variables

Model-7	MOdel-8	model-9
3.306	5.503	3.685

1.3 Linear model for three variables

Table 4: MSE for linear regression(interacting) model with three variables

Model-10	MOdel-11
3.949	3.212

1.4 Anova for all the linear model

```
## Analysis of Variance Table
##
## Model  1: Omni.LLR.int ~ LLR.x
## Model  2: Omni.LLR.int ~ LLR.y
## Model  3: Omni.LLR.int ~ LLR.z
## Model  4: Omni.LLR.int ~ LLR.x + LLR.y
## Model  5: Omni.LLR.int ~ LLR.y + LLR.z
## Model  6: Omni.LLR.int ~ LLR.z + LLR.x
## Model  7: Omni.LLR.int ~ LLR.x + LLR.y + LLR.y:LLR.x
## Model  8: Omni.LLR.int ~ LLR.y + LLR.z + LLR.z:LLR.y
## Model  9: Omni.LLR.int ~ LLR.z + LLR.x + LLR.x:LLR.z
## Model 10: Omni.LLR.int ~ LLR.z + LLR.y + LLR.x
## Model 11: Omni.LLR.int ~ LLR.x + LLR.y + LLR.z + LLR.x * LLR.y * LLR.z
##    Res.Df    RSS Df Sum of Sq  Pr(>Chi)
## 1     818 4157.3
## 2     818 6746.9  0  -2589.51
## 3     818 6039.9  0    706.93
## 4     817 3268.6  1   2771.33 < 2.2e-16 ***
## 5     817 5352.5  0  -2083.87
```

```
## 6     817 3789.3  0   1563.12
## 7     816 2710.5  1   1078.83 < 2.2e-16 ***
## 8     816 4512.1  0  -1801.56
## 9     816 3021.8  0   1490.26
## 10    816 3237.9  0   -216.06
## 11    812 2634.2  4    603.68 < 2.2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

From the linear regression, The MSE of the model varies from 6.527 being the highest and 3.212 being the lowest. The model 11 that interact with all the variable has the least MSE. Likewise, from the ANOVA test, model 4, model 7 and model 11 seem to have statistically significant. ANOVA shows that the variable x, y has a linear relationship. each model looking significant has the variable either x and y or the all 3 variables. Hence for linear regression we can say that x and y have linear relationship also we can say that x, y and z have linear relationship.

We cannot be sure since visually the data looks some what quadratic therefore I would like to fit the quadratic model and varify.

2. Quadratic model

2.1 Quadratic model for single variable

Table 5: MSE for quadratic regression model with single variable

Model-1	MOdel-2	model-3
5.07	5.07	7.366

2.2 Quadratic model for two variables

Table 6: MSE for quadratic regression model with two variables

Model-4	MOdel-5	model-6
2.883	5.033	2.378

2.3 Quadratic model for three variables

```
## [1] "R squared quadratic model for model 7 is : 0.786098674276554"
```

```
## [1] "MSE of quadratic model for model 7is : 2.78421138468662"
```

2.4 Anova for quadratic models

```
## Analysis of Variance Table
##
## Model 1: Omni.LLR.int ~ LLR.x + llrx2
## Model 2: Omni.LLR.int ~ LLR.y + llry2
## Model 3: Omni.LLR.int ~ LLR.z + llrz2
## Model 4: Omni.LLR.int ~ LLR.x + LLR.y + llry2 + LLR.x:(LLR.y + llry2)
## Model 5: Omni.LLR.int ~ LLR.y + LLR.z + llrz2 + LLR.y:(LLR.z + llrz2)
## Model 6: Omni.LLR.int ~ LLR.z + LLR.x + llrx2 + LLR.z:(LLR.x + llrx2)
## Model 7: Omni.LLR.int ~ LLR.x + LLR.y + LLR.z + LLR.x:(LLR.y + llry2) +
```

```
##     LLR.x:(LLR.z + llrz2)
##   Res.Df    RSS Df Sum of Sq  Pr(>Chi)
## 1    817 2787.5
## 2    817 5983.0  0   -3195.5
## 3    817 5333.8  0     649.3
## 4    814 2364.1  3    2969.6 < 2.2e-16 ***
## 5    814 4127.1  0   -1762.9
## 6    814 1949.7  0    2177.3
## 7    812 2283.1  2    -333.3
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Since, the same process is repeated in the quadratic model. MSE of the models is viewed. The MSE of the models varies from 7.36 being maximum and 2.3 being the minimum model-6 that has the x and y interacting with the other models shows the least MSE. The ANOVA analysis shows only one model as significant. We did not get any values for the other models. That may be because of the values that are too small to calculate. However, model 6 appears to be significant.

Again, I would like to compare the MSE for both linear and quadratic models. Since the MSE for the linear regression that I assumed better fit has the value of 3.212 and the better fit for the quadratic model is 2.3 therefore because of the least MSE, the quadratic model has the better fit with the interaction between x and y. The model that we find to have the better is `Omni.LLR.int ~ LLR.x + LLR.y + llry2 + LLR.x:(LLR.y + llry2)` Therefore LLR.x and LLR.y interaction with x has the better fit and has the quadratic relationship.

Hence, from the linear and the quadratic model, it apperas that x and y have linear realationship.The same thing can be proved for the quadratic model, however the quadratic model is the better fit. Therefore, from my analysis, I find out that the x and y has a quadratic relation.