Question 2

Amin Baabol

11/28/2020

##Introduction

We are provided with an excel file containing 820 samples. In this files there are log-likelihood ratios stored in four variables (LLRX,LLRY,LLRZ,Omnibus). Our goal for this analysis is aimed at establishing whether or not a relationship between variable “Omnibus” and the remaining three variables exists.

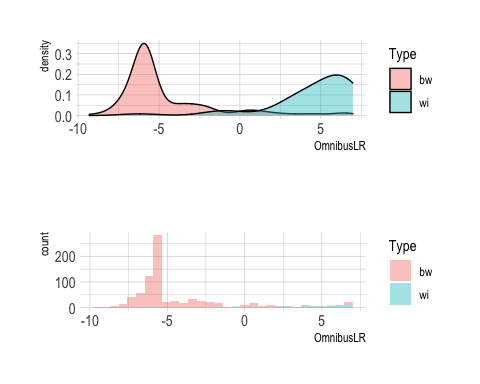
## Approach

In order for our analysis to fully capture the possibility of a relationship existing between the variables, we will conduct an unpaired t-test to confirm that difference in the means exists. In doing so, the original data will be split by type and compare the two types’ means. We will then visualize the distributions of data.Moreover,performing an F-test in one way nalysis of variance will should confirm our unpaired t-test results.The Equation needed to compute this F-test is: We will not be comparing numerous models, we are simply constructing a few linear models and a general additive model to characterizes the influence of LLRX,LLRY, LLRZ on OmnibusLLR.

##Mean Difference

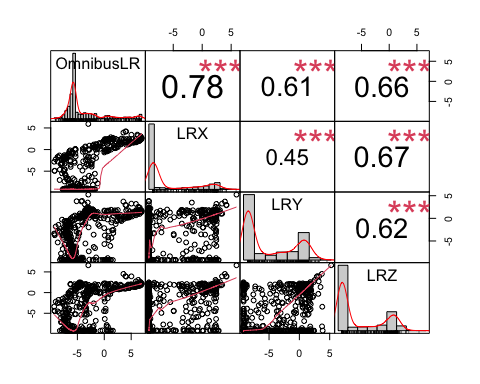
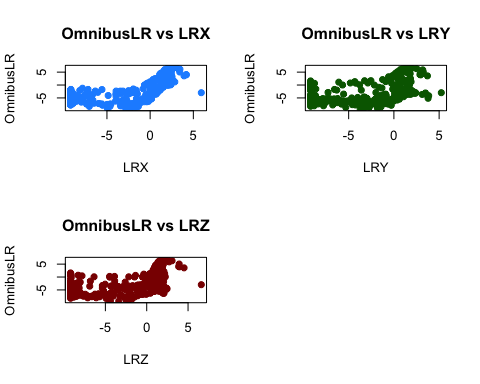
Performing Welch two Sample t-test we find that the p-value is significant enough to reject the null hypothesis that the two means not are different,which suggests that the presence of the LLRX,LLRY,LLRZ have influence in the mean difference of OmnibusLLR in the two groups.we will simply compare the Omnibus density plots for the two groups, “bw” and “wi”.The density and histogram plots also indicate the unequal distribution of Omnibus for the two types.

##   
## Welch Two Sample t-test  
##   
## data: wi$OmnibusLR and bw$OmnibusLR  
## t = 20.345, df = 43.986, p-value < 2.2e-16  
## alternative hypothesis: true difference in means is not equal to 0  
## 95 percent confidence interval:  
## 8.245494 10.058674  
## sample estimates:  
## mean of x mean of y   
## 4.576042 -4.576042



##Correlation plot and Scatterplots

The correlation plot indicates that variable “OmnibusLLR” is correlated with LLRX at a rate of 78%, while the same variable is correlated with LLRY and LLRZ at a rate of 61% and 66% respectively. LLRX,LLRY and LLRZ have low correlations among themselves which means the risk of multicollinearity is not eminent.Furthermore, the scatter plots indicate intricate relationships between Omnibus LLR and the other log-likelihood ratio variables.This complex pattern can not simply be explained by a linear regression model.



## Linear Model Assumptions

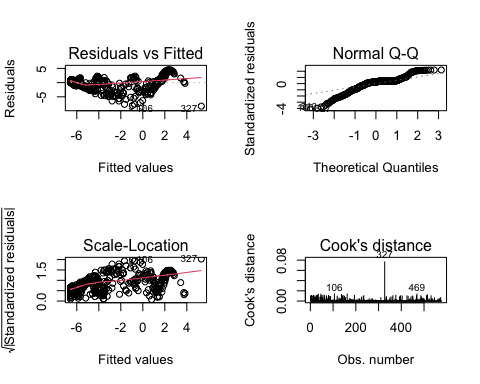
Linear models make quite a few assumptions but the most important ones include, a linear relationship exists between the response variable “OmnibusLLR” and the covariates.The residuals are normally distributed at around zero within reasonable standard deviation.It also assumes that the variance in the residuals is constant.

## Linear regression Model

A total of six linear models were constructed.A one way analysis of variance was conducted, comparing both their F-statistics and Residuals. A likelihood ratio test using *anova()* was also performed. It turns out the model with all three predictors, LLRX,LLRY, and LLRZ performed best. All predictors were significant at the 0.05 level. However,the residuals failed to hold up the normal distribution assumption. Which means linear models cannot satisfactorily explain this relationship.

## Analysis of Variance Table  
##   
## Model 1: OmnibusLR ~ 1  
## Model 2: OmnibusLR ~ LRX  
## Model 3: OmnibusLR ~ LRY  
## Model 4: OmnibusLR ~ LRZ  
## Model 5: OmnibusLR ~ LRX + LRY + LRZ  
## Model 6: OmnibusLR ~ LRX + I(LRY^2) + I(LRZ^3)  
## Res.Df RSS Df Sum of Sq Pr(>Chi)   
## 1 575 7938.0   
## 2 574 3037.0 1 4901.0 < 2.2e-16 \*\*\*  
## 3 574 4876.7 0 -1839.7   
## 4 574 4411.8 0 464.9   
## 5 572 2366.0 2 2045.8 < 2.2e-16 \*\*\*  
## 6 572 2672.1 0 -306.1   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##   
## Call:  
## lm(formula = OmnibusLR ~ LRX + LRY + LRZ, data = train.data)  
##   
## Residuals:  
## Min 1Q Median 3Q Max   
## -8.2841 -0.7167 0.5220 0.7435 4.4653   
##   
## Coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) 0.73300 0.15526 4.721 2.95e-06 \*\*\*  
## LRX 0.49205 0.02544 19.344 < 2e-16 \*\*\*  
## LRY 0.23096 0.02414 9.567 < 2e-16 \*\*\*  
## LRZ 0.06940 0.02834 2.449 0.0146 \*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Residual standard error: 2.034 on 572 degrees of freedom  
## Multiple R-squared: 0.7019, Adjusted R-squared: 0.7004   
## F-statistic: 449 on 3 and 572 DF, p-value: < 2.2e-16



##Generalized Additive Models

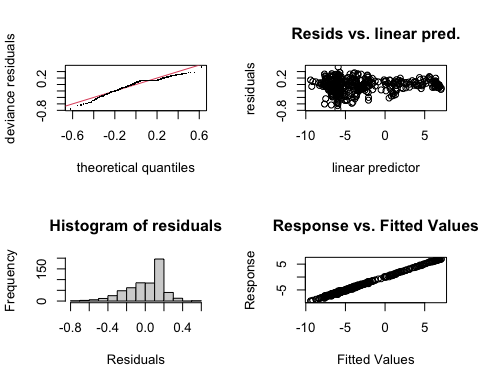
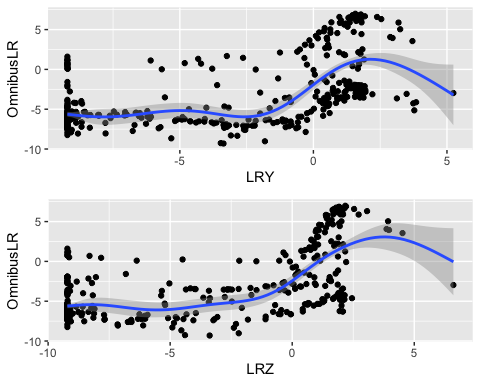
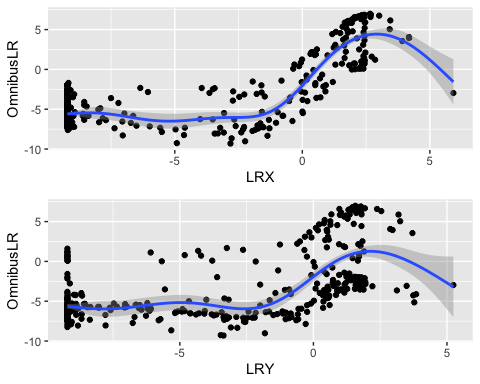
We are approaching this model as a non-parametric model, which means we are not making too many assumptions.We will estimate the appropriate functional form of the relationship from the data.

##   
## Family: gaussian   
## Link function: identity   
##   
## Formula:  
## OmnibusLR ~ s(LRX) + s(LRY) + s(LRZ)  
##   
## Parametric coefficients:  
## Estimate Std. Error t value Pr(>|t|)   
## (Intercept) -4.03941 0.00824 -490.2 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## Approximate significance of smooth terms:  
## edf Ref.df F p-value   
## s(LRX) 8.696 8.950 6498.2 <2e-16 \*\*\*  
## s(LRY) 9.000 9.000 2471.1 <2e-16 \*\*\*  
## s(LRZ) 8.709 8.947 667.6 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## R-sq.(adj) = 0.997 Deviance explained = 99.7%  
## GCV = 0.041064 Scale est. = 0.03911 n = 576

## (Intercept) s(LRX).1 s(LRX).2 s(LRX).3 s(LRX).4 s(LRX).5   
## -4.0394153 1.6666138 31.2595530 1.8244183 -10.5374601 3.6784339   
## s(LRX).6 s(LRX).7 s(LRX).8 s(LRX).9 s(LRY).1 s(LRY).2   
## 8.5742690 -3.9949033 23.3119075 -4.6158101 -0.3511293 -3.3778096   
## s(LRY).3 s(LRY).4 s(LRY).5 s(LRY).6 s(LRY).7 s(LRY).8   
## -1.7490051 -4.3519703 1.7101122 6.2494993 -0.1123141 -7.3144193   
## s(LRY).9 s(LRZ).1 s(LRZ).2 s(LRZ).3 s(LRZ).4 s(LRZ).5   
## 3.9627172 -5.2518988 -13.2118808 -1.0249758 6.1441922 -3.7072000   
## s(LRZ).6 s(LRZ).7 s(LRZ).8 s(LRZ).9   
## -6.2603988 -1.5145042 16.6561141 -5.6650334

## [1] 21.4557

## [1] -203.6793



##   
## Method: GCV Optimizer: magic  
## Smoothing parameter selection converged after 20 iterations.  
## The RMS GCV score gradient at convergence was 8.015429e-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(LRX) 9.00 8.70 0.97 0.23   
## s(LRY) 9.00 9.00 0.72 <2e-16 \*\*\*  
## s(LRZ) 9.00 8.71 0.50 <2e-16 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

## RMSE R2  
## 1 0.1747169 0.9972704

## Conclusion

The general additive model has shown to be capable of characterizing the relationship between the OmbibusLLR and the three covariates LLRX,LLRY and LLRZ. This complex non-parametric relationship cannot be adequately explained by a simple linear regression. The linear regressions that was constructed failed to hold up a few of the assumptions. The residuals in particular were not normally distributed in the qq-normal plot. The linear regression’s predictions only seemed to be accurately predicting perhaps the first few iterations at which point it loses reliability.On the other hand, the general additive model’s prediction seems to capture well the pattern in the observed data.This means given the low MSE,AIC and Deviance the gam model shows in the summary,its predictive ability is superior to that of the linear regression. However,the main reason this gam model was constructed is to examine the effect or the influence the covariates have on the response variable “OmnibusLLR”. According to the summary statistics of the gam model, LLRX,LLRY, and LLRZ have no significance as parametric terms. It is only when they’re smoothed that they become very significant with p-values of less than 2^-16. This indicates LLRX,LLRY,and LLRZ have significant non-linear influence or effect on the response variable OmnibusLLR.Moving forward, I would recommend performing a polynomial regression with varying degrees to characterize the

##Works Cited

Michael, Semhar, and Christopher P. Saunders. “Scatterplot Smoothers and GAM” Chapter 10. 05 DEc. 2020, South Dakota State University, South Dakota State University.

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

ackson, Simon. “Visualising Residuals • BlogR.” BlogR on Svbtle, drsimonj.svbtle.com/visualising-residuals.

Lowhorn, J. (n.d.). Retrieved DEcember 05, 2020, from <https://rstudio-pubs-static.s3.amazonaws.com/326465_9748350bbfca41afb753211eff074761.html>

Neupane, Achal. “STAT\_601\_Final.” Achal Neupane, 04 Dec. 2019, achalneupane.github.io/achalneupane.github.io/post/stat\_601\_final/.

Priyadarshana, Pandula. “Sign In.” RPubs, rpubs.com/PandulaP/logisticregression\_model\_compare.

Kassambara, kassambara, et al. “Linear Regression Assumptions and Diagnostics in R: Essentials.” STHDA, 11 Mar. 2018, www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/.