Amin Baabol
12/02/2020

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: $F = Variance - Between - Variables / Variance - Within - Variables$ 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 at 2.2e-16 to reject the null hypothesis that the two means are not 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.

Correlation

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 cannot 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.

Generalized Additive Model

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.

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 were 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.

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

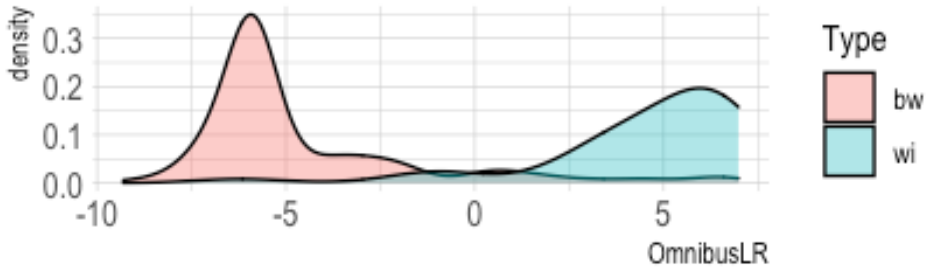
Lowhorn, J. (n.d.). Retrieved DEcember 05, 2020, from https://rstudio-pubs-static.s3.amazonaws.com/326465_9748350bbfca41afb753211eff074761.html

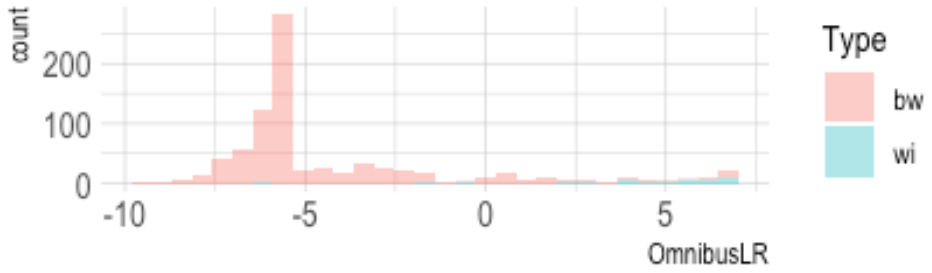
Neupane, Achal. “STAT_601_Final.” Achal Neupane, 04 Dec. 2019, achalneupane.github.io/achalneupane.github.io/post/stat_601_final/.

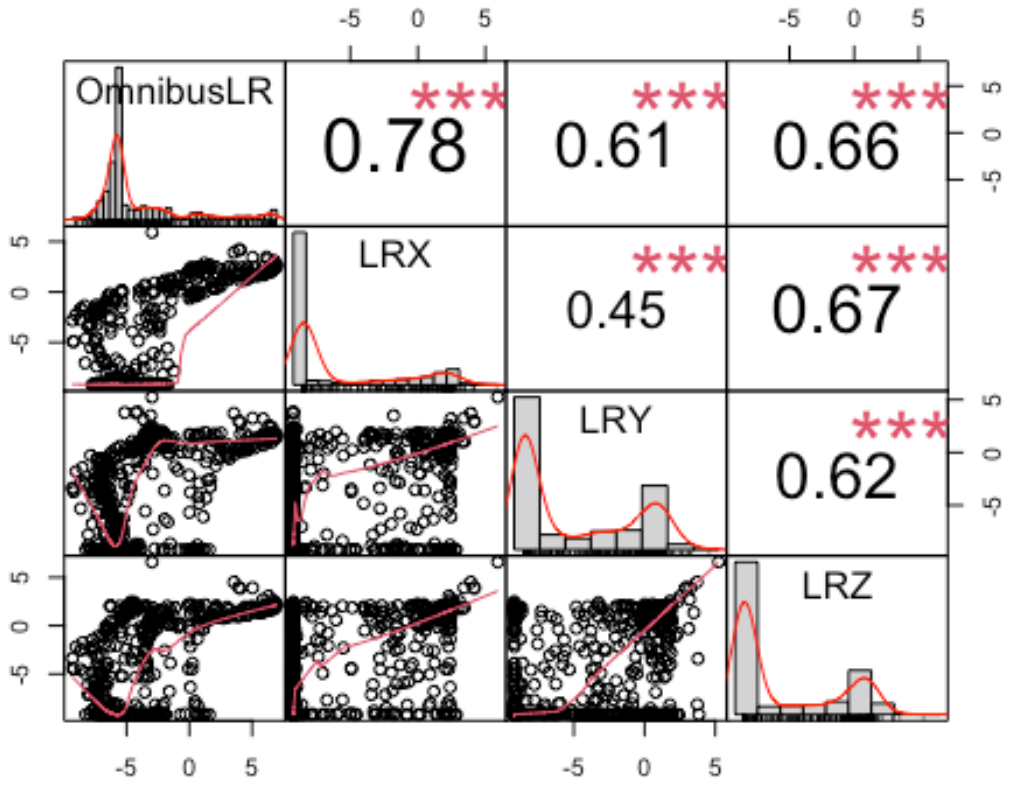
Priyadarshana, Pandula. “Sign In.” RPubs, rpubs.com/PandulaP/logisticregression_model_compare.

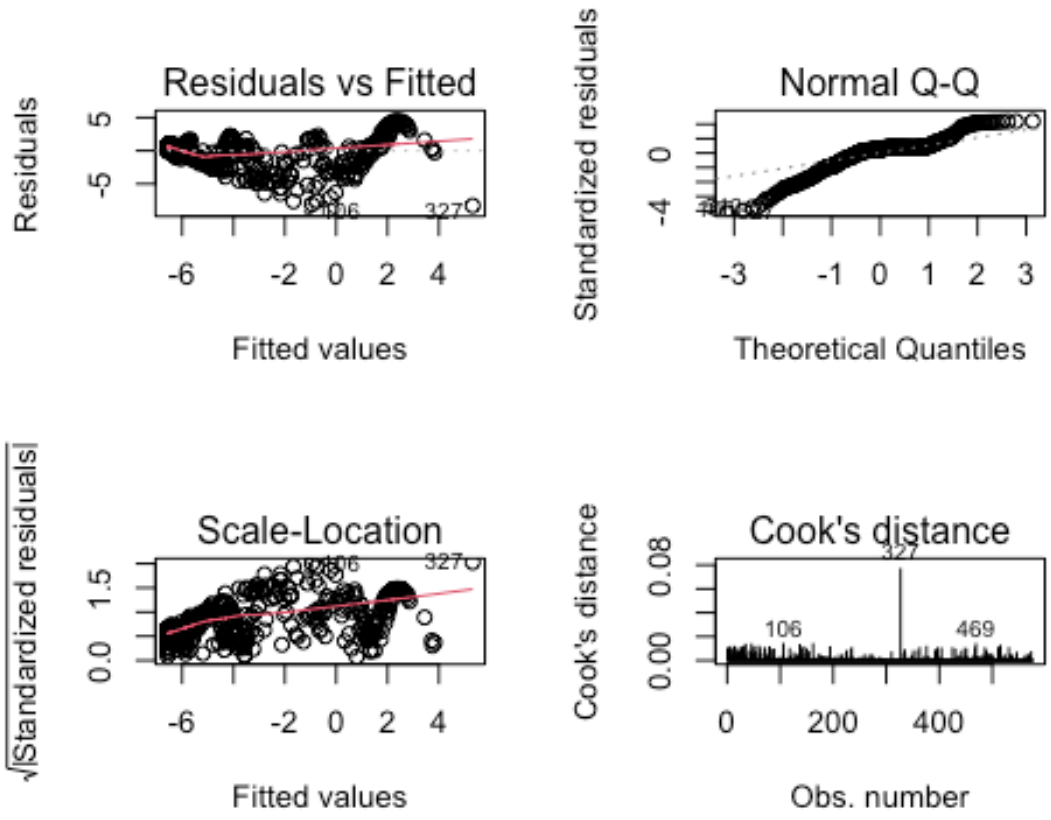
Kassambara, kassambara, et al. “Linear Regression Assumptions and Diagnostics in R: Essentials.” STHDA, 11 Mar. 2018, www.sthda.com/english/articles/39-regression-model-diagnostics/161-linear-regression-assumptions-and-diagnostics-in-r-essentials/.

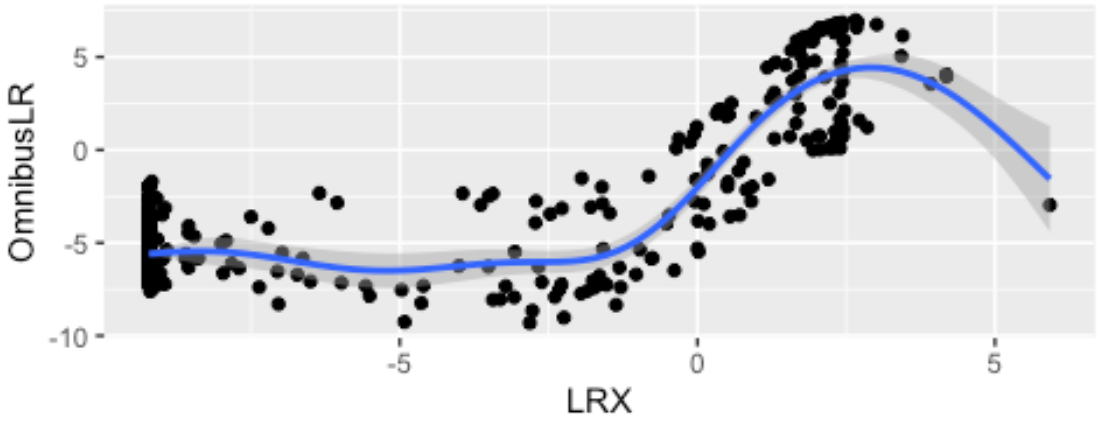
Appendix

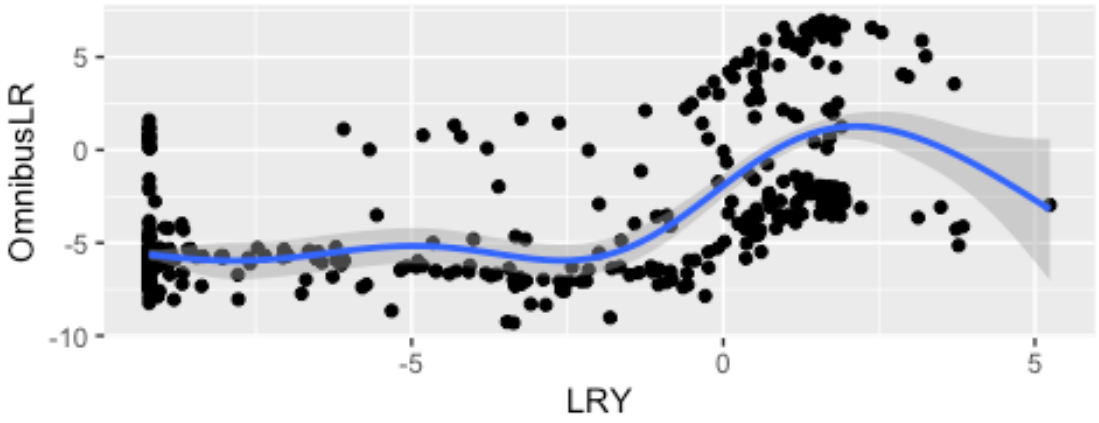

density
0.3
0.2
0.1
0.0
-10
-5
0
5
OmnibusLR
Type
bw
wi

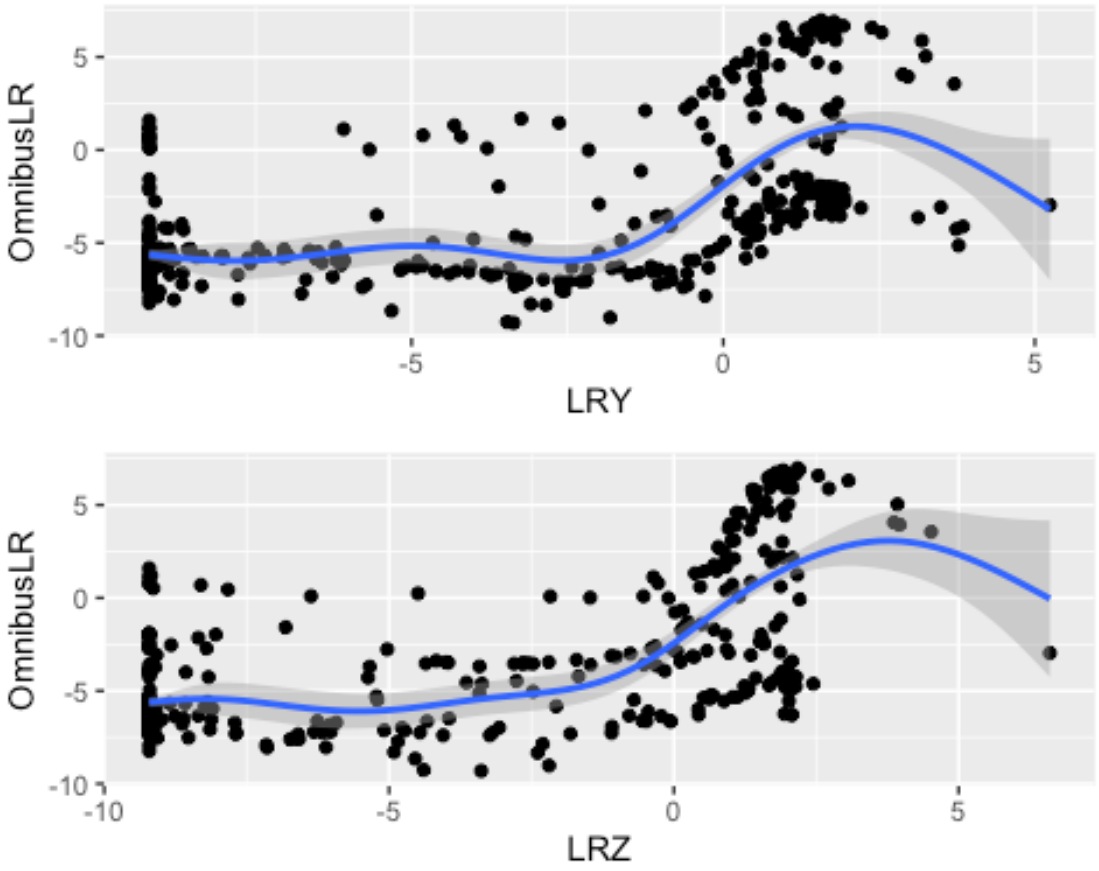

count
200
100
0
-10
-5
0
5
OmnibusLR
Type
bw
wi

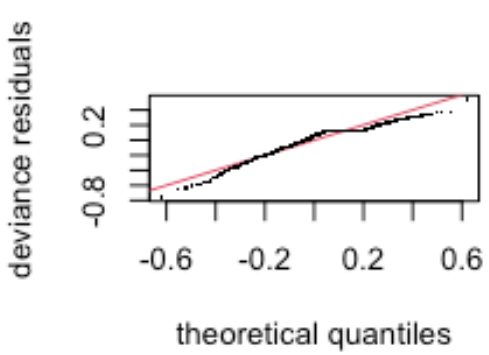

OmnibusLR
LRX
LRY
LRZ
0.78 ***
0.61 ***
0.66 ***
0.45 ***
0.67 ***
0.62 ***
-5 0 5

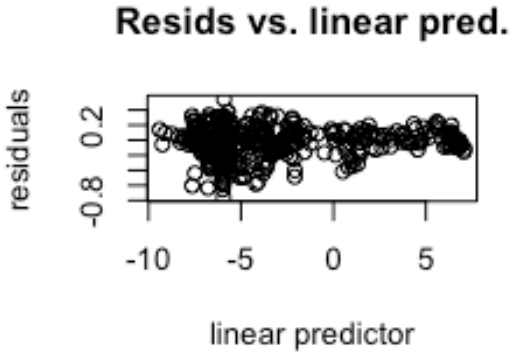

Residuals vs Fitted
Residuals
Fitted values
Normal Q-Q
Standardized residuals
Theoretical Quantiles
Scale-Location
√|Standardized residuals|
Fitted values
Cook's distance
Cook's distance
Obs. number
106
327
469

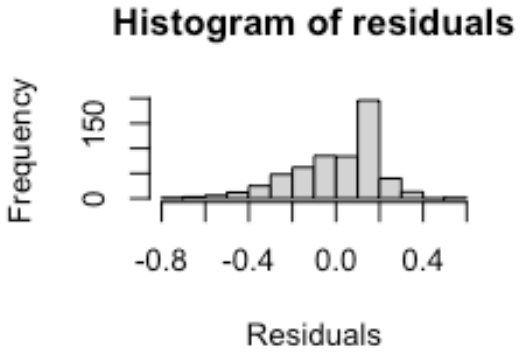

OmnibusLR
5
0
-5
-10
-5
0
5
LRX

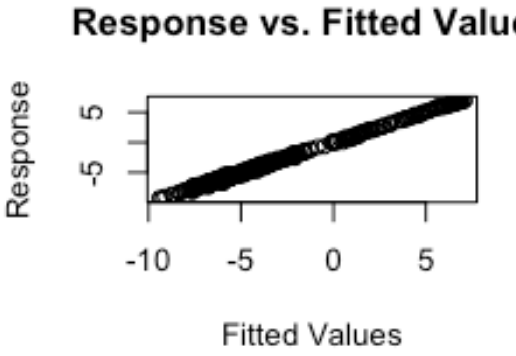

OmnibusLR
5
0
-5
-10
-5
0
5
LRY


OmnibusLR
5
0
-5
-10
-5
0
5
LRY
OmnibusLR
5
0
-5
-10
-10
-5
0
5
LRZ


deviance residuals
-0.8
0.2
-0.6
-0.2
0.2
0.6
theoretical quantiles


Resids vs. linear pred.
residuals
-0.8
0.2
-10
-5
0
5
linear predictor


Histogram of residuals
Frequency
0
150
-0.8
-0.4
0.0
0.4
Residuals


Response vs. Fitted Values
Response
-5
5
-10
-5
0
5
Fitted Values