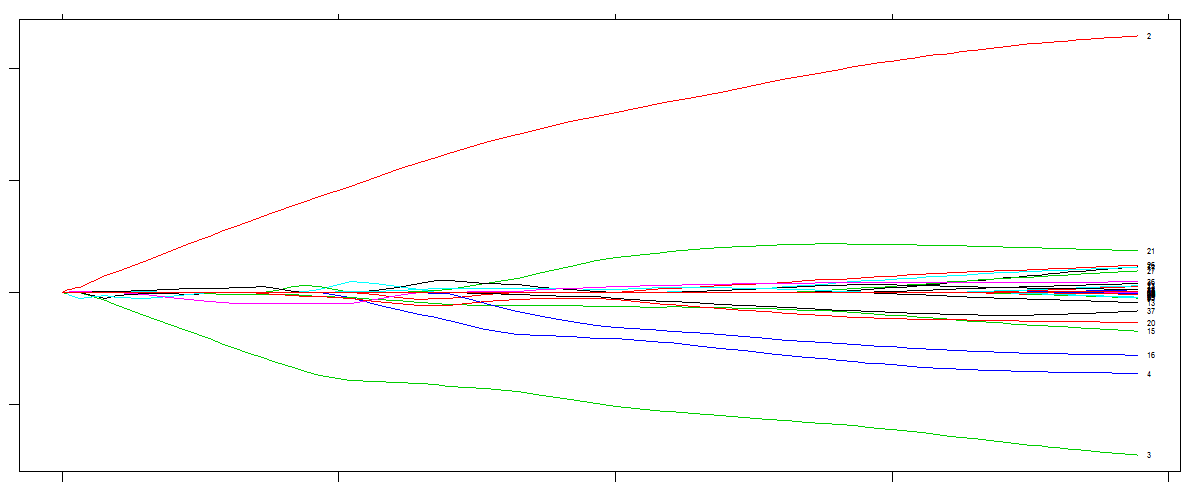
**Kalbi Zongo ST 599 Project3 Report   
----------------------------------------------------------------------**

**1. Project /data background**

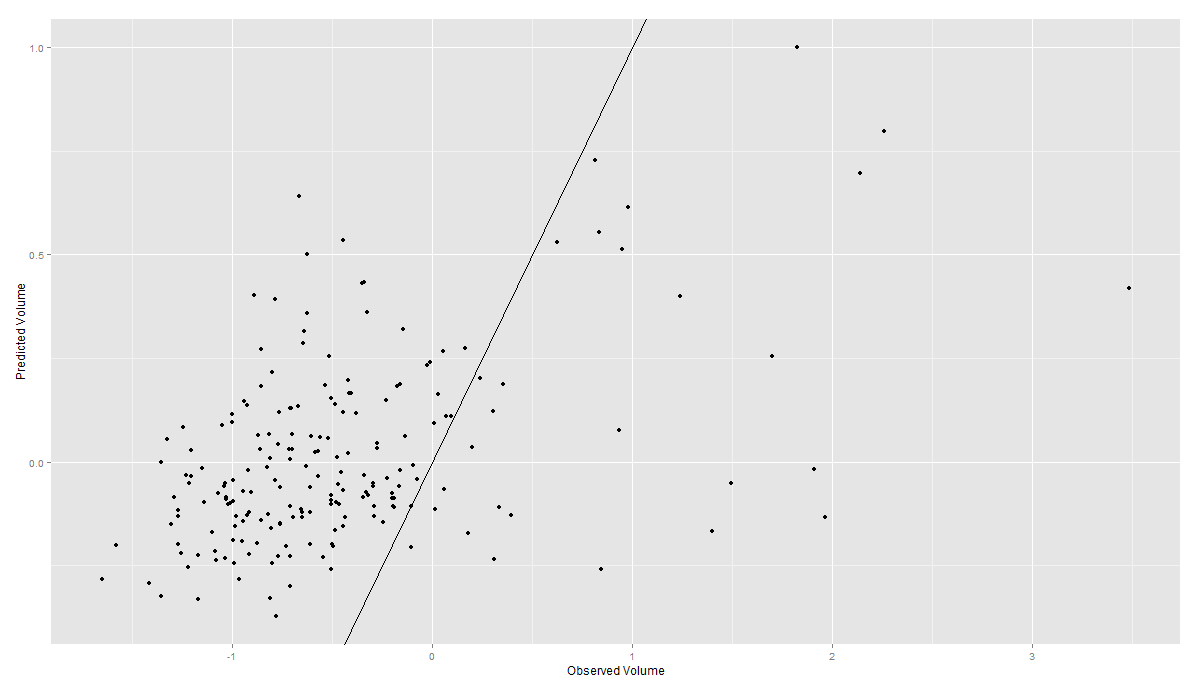
QuantQuote is a leading provider of high resolution historical intraday stock data and live feeds. The database contains stock prices of around five hundred (500) companies from 1998 to present. The datasets contain variable such as the date, time, open prices, high prices, low prices, closed prices, and volumes-trading volume during an interval. The goal of this project is to predict the stocks trading volume given stocks prices. In other words, what is the stock trading volume for a future date using dataset from table\_a?

**2.2 Method and results**

The data consists of 3452 observations and 7 variables. The time variable has zero values for all observations, so I removed that variable and the date variable from analysis. Data were split into training set (2000 Obs) and test set (1452 Obs). A multiple regression model was poorly fitting the data (R-sq-adjusted =.07%). I fitted 8 degrees polynomial regression with only second order interaction terms. This has improve the proportion of variation explain by the model by 21%. I fitted a Ridge and Lasso model to the data with 8 degrees polynomial terms. While the average prediction error is almost the same for multiple regression and Ridge regression (1.720 and 1.719 respectively), Lasso has the smallest average prediction error of .816. the lasso coefficients are plotted below against L1 Norm (a function of the tuning parameter). The open price (V2) and high price (V3) seem to be very important regardless of penalized value (lambda). As lambda get smaller, much coefficients become non-zero.



The scatter plot of observed volume vs. predicted volume below reveals that predicted volume of stocks are much higher than the actually observed volumes. The dots should have fall around and close to the line with slope one (1). More predictors are needed to improve the predictions.



**3. Obstacles**

Number of variable mention in documentations were not found in datasets-these were earning values, split factor, dividend. These predictors could have help to improve the stocks volume predictions. To prevent high bias due to fewer predictors (p) for too large N, I fitted a 8 degrees polynomial model with only first interaction terms. Future work include collecting more predictors associated with stock volumes, taking into account time and date of the stocks. The project github directory will be provided by email by Friday 5.

**References**

<https://quantquote.com/support_documentation.php>