**FE-582 Foundations of Financial Data Science, Spring 2021**

**Group 5: Daniel Bachalis, Gurjivan Kalkat, Naveen Nagarajan**

**Forecasting stock prices during pandemic**

**Introduction:**

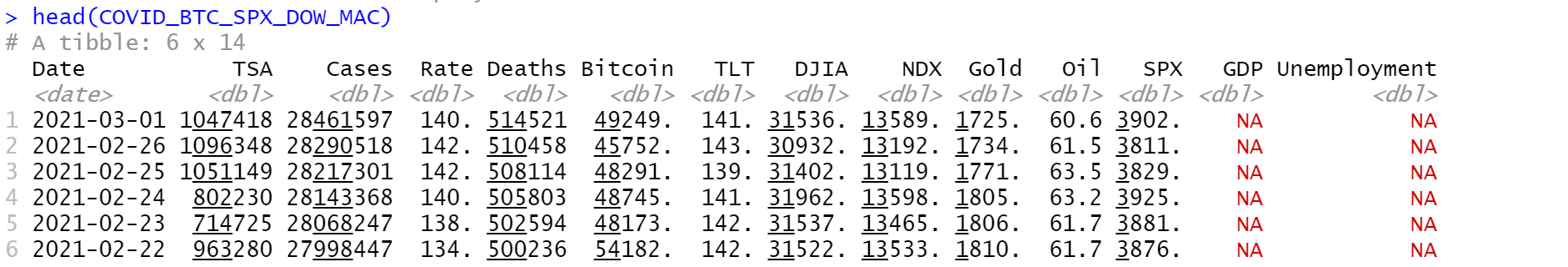
The current pandemic has affected global markets and economies tremendously, especially in the United States. The research being performed will look at the performance and correlations across commodities, macroeconomic indicators, equity indices, COVID 19 data, and TSA data before and after the pandemic. Certain assets and commodities performed better post pandemic while others are still lagging in terms of returns. An initial model will be created starting from the beginning of 2019 to the start of the pandemic. The model will then be changed to account for the current environment. Based upon the analysis of this data, a determination will be made as to the best asset allocation for investors and what can be expected going forward.

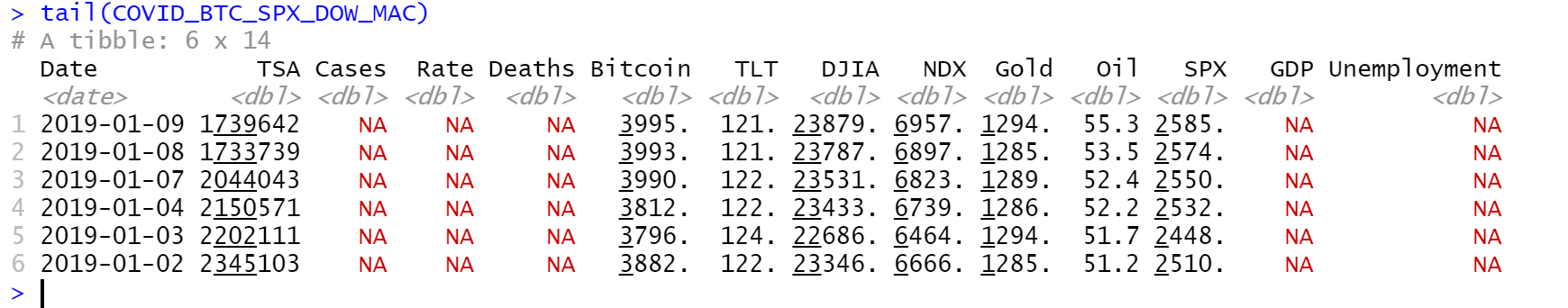
**Research statement (questions):**

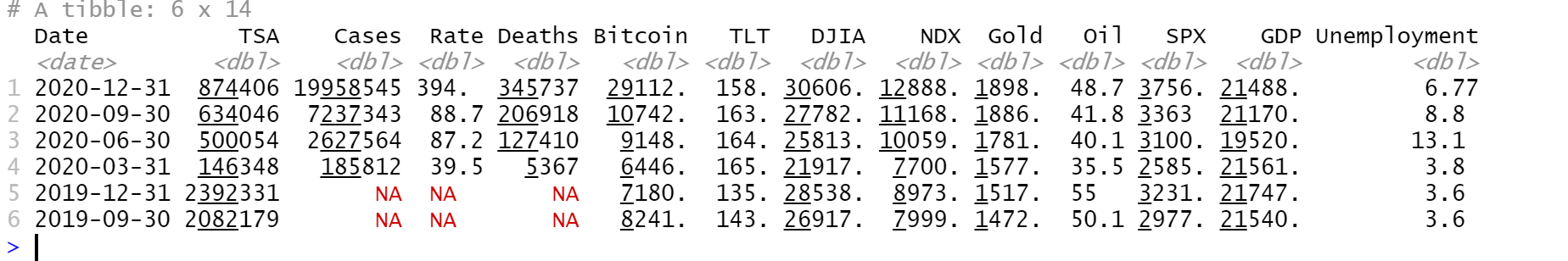
We will use data science techniques to do the following:

* Predict which business cycle we are currently in and how long it will last as it relates to the Covid-19 pandemic.
  + Develop Covid-19 prediction model to leverage for pandemic duration prediction.
  + Compare Data beginning from the pre pandemic levels starting from 2019 until March 2021, to observe how different asset classes and macroeconomic data has been affected by Covid-19
* Make recommendations for portfolios and asset allocation in order to maximize returns and hedge against risk.

**Data:**

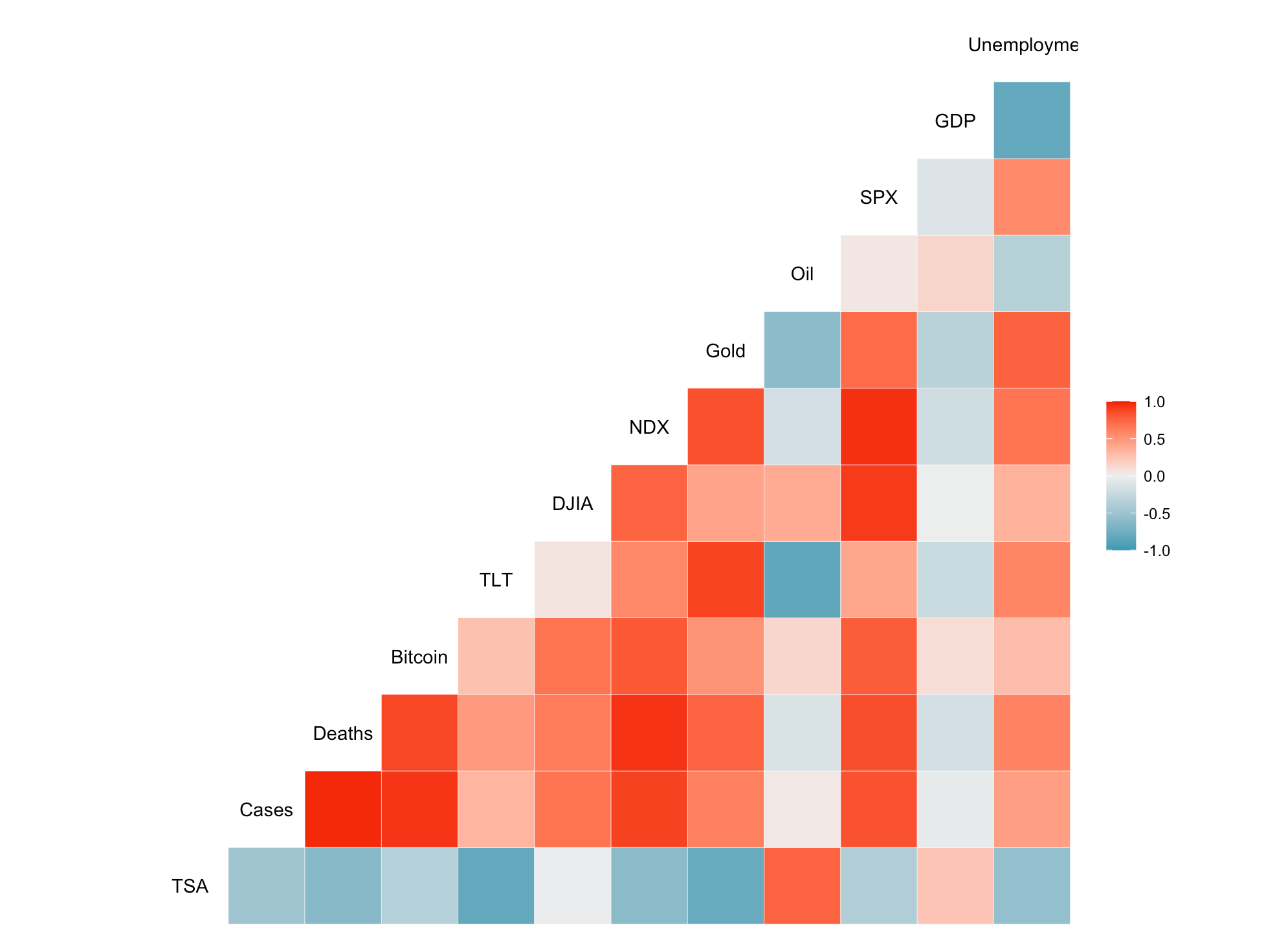






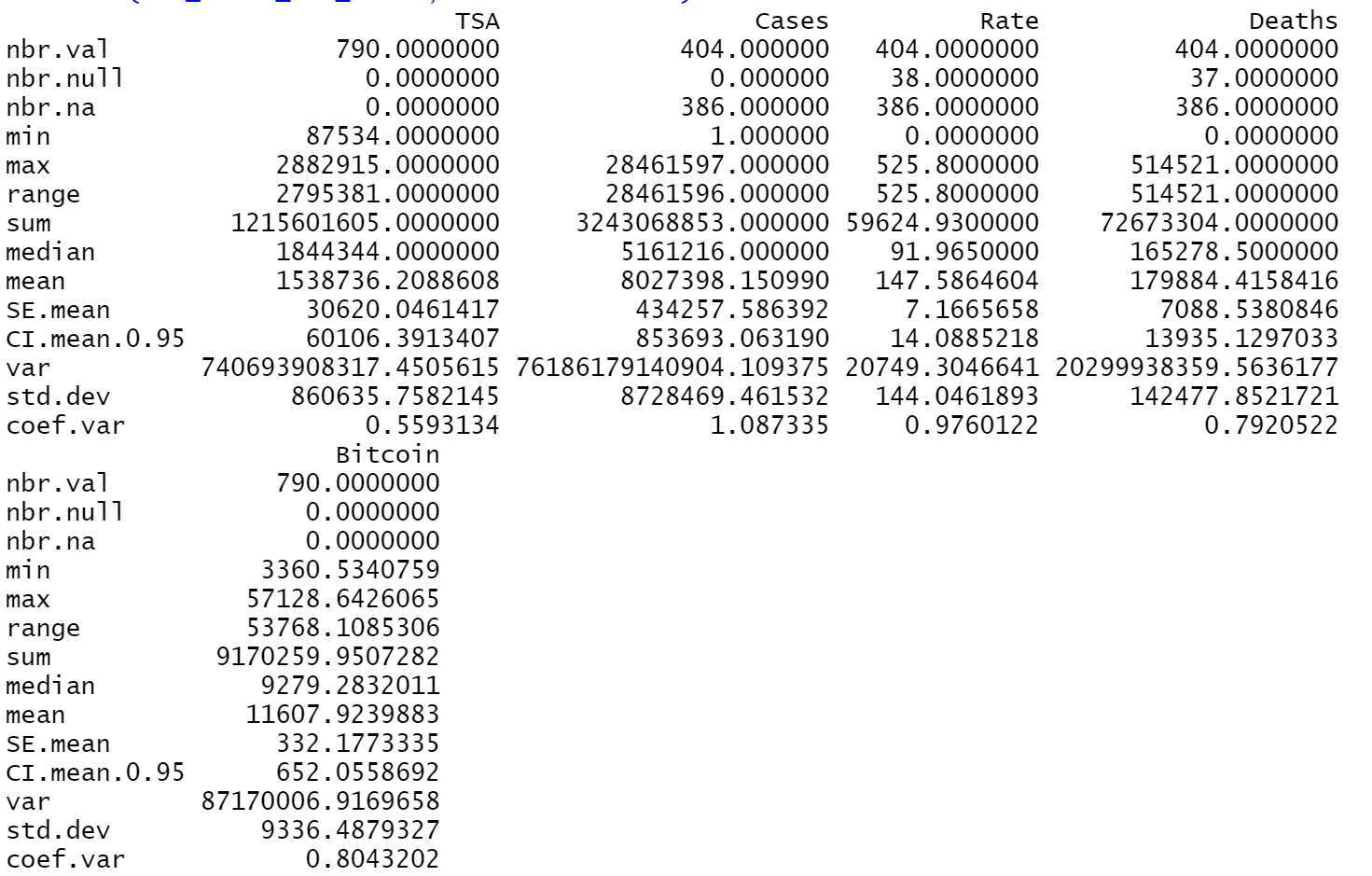
Datasets were retrieved from bloomberg [3]

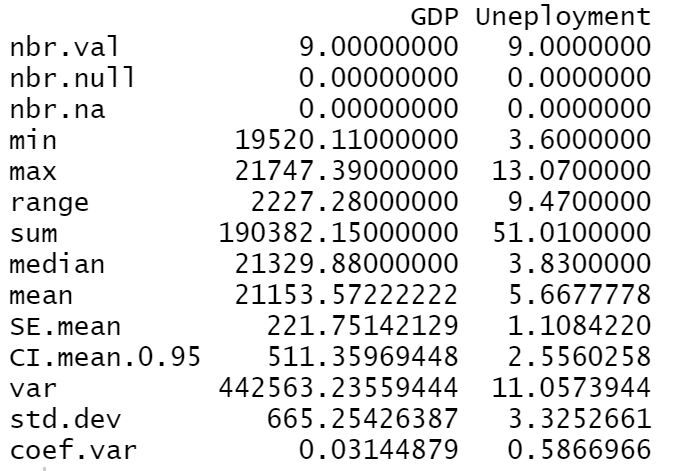
Correlation matrix



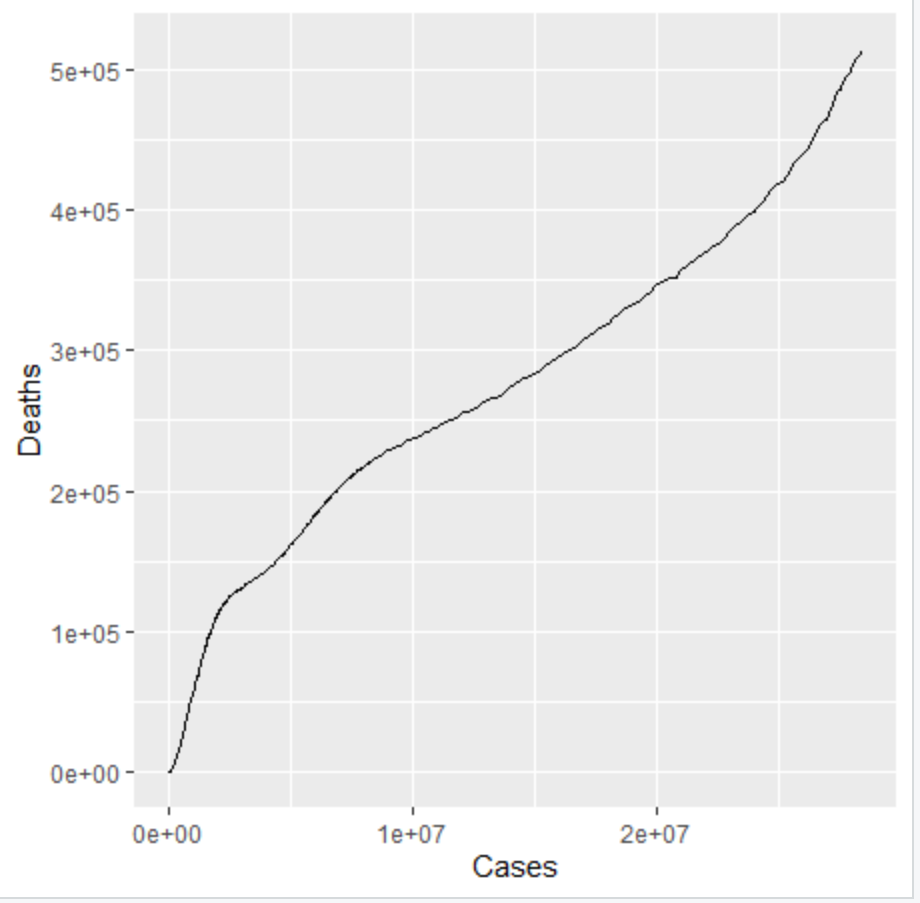
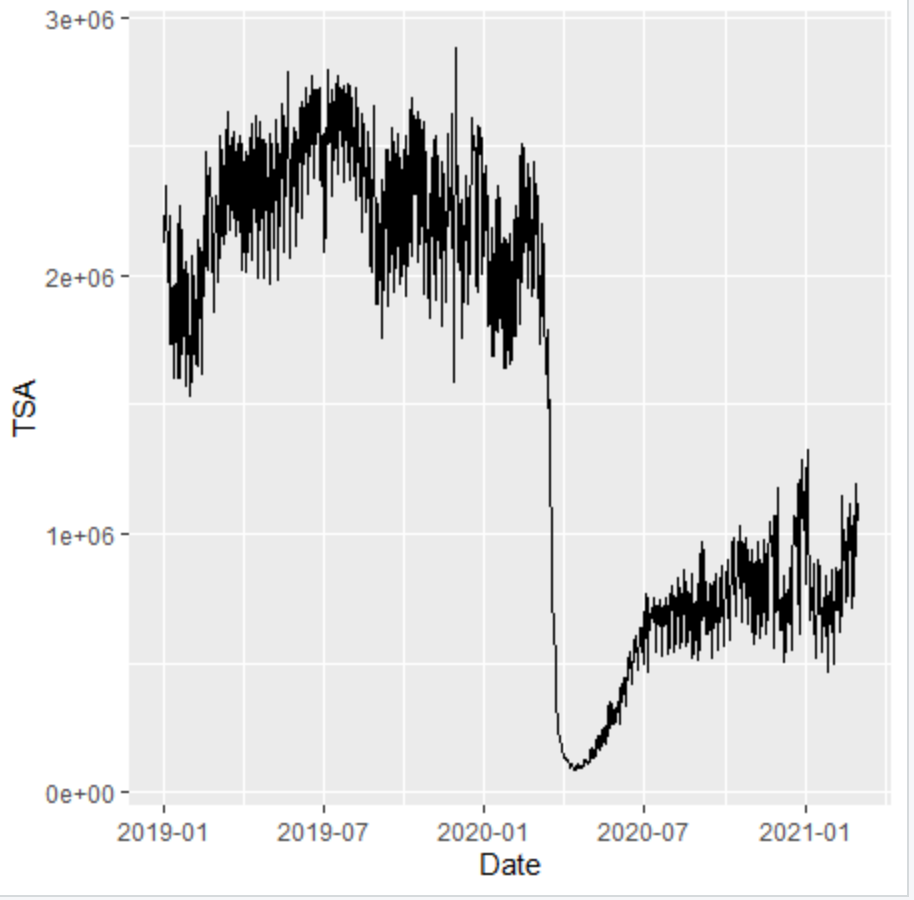
**Exploratory Data Analysis (EDA):**

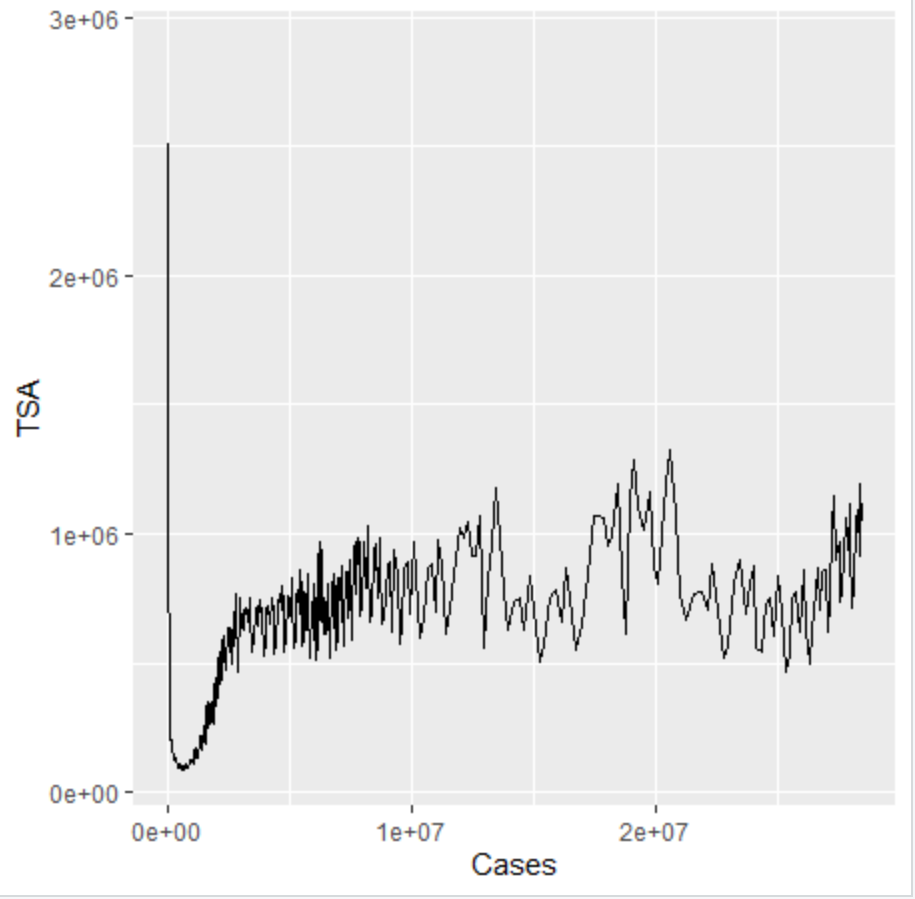
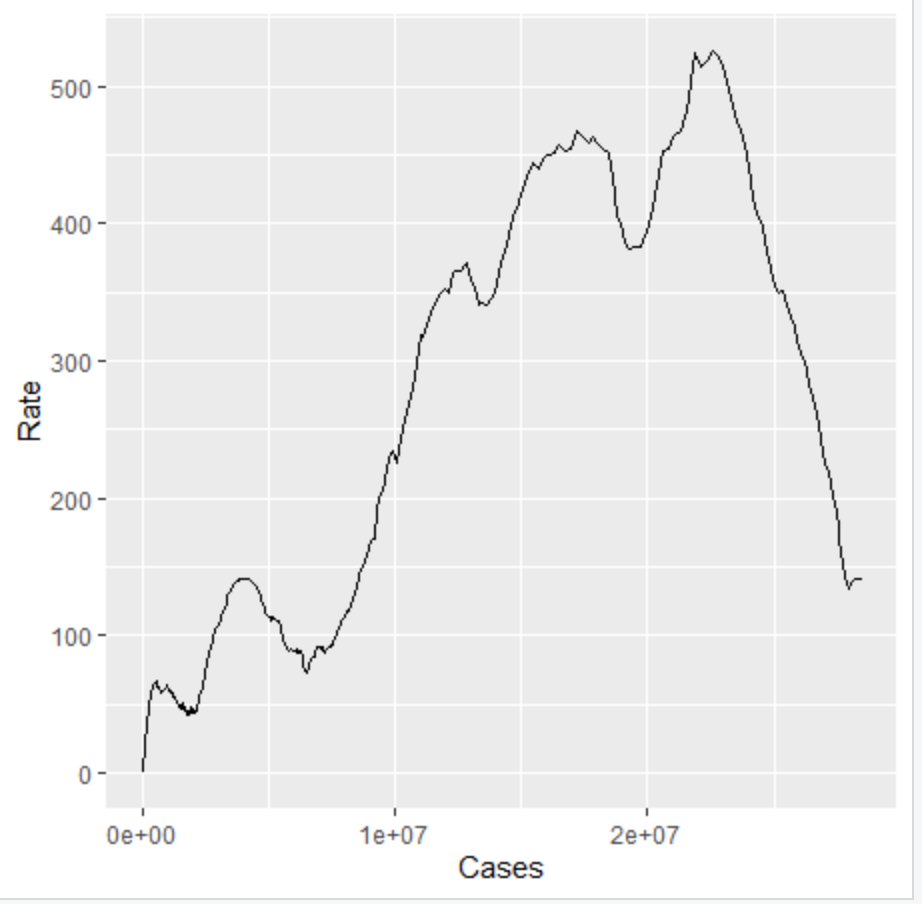
The following figures below display the initial summary statistics for all of the data:





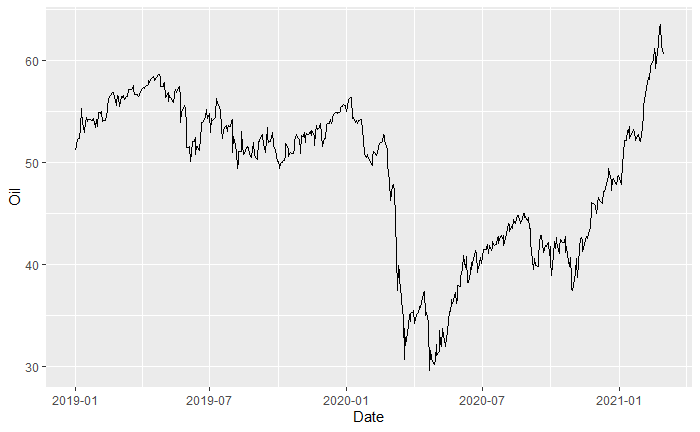
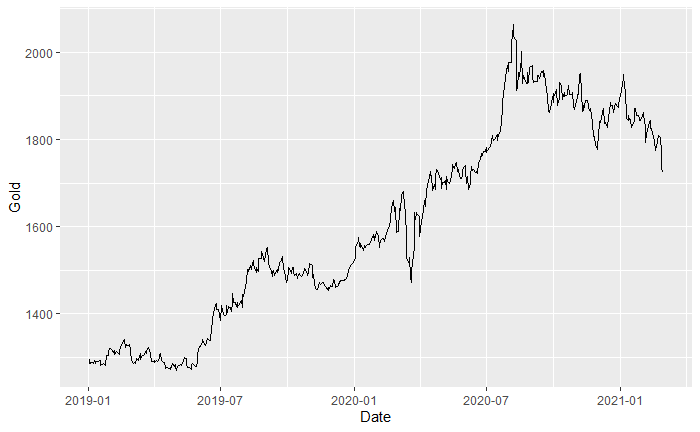
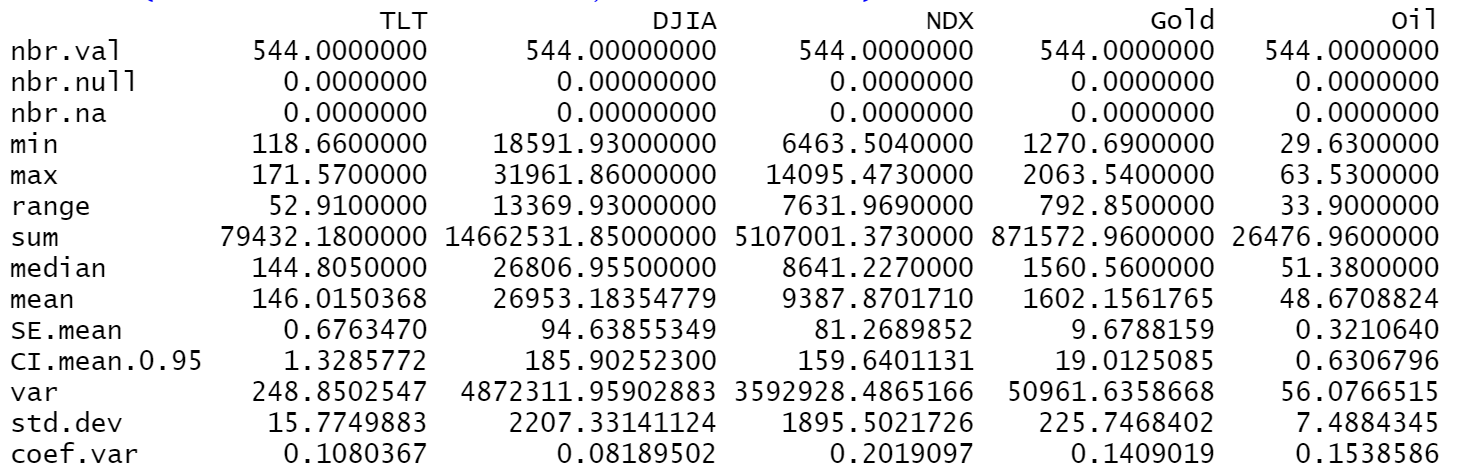
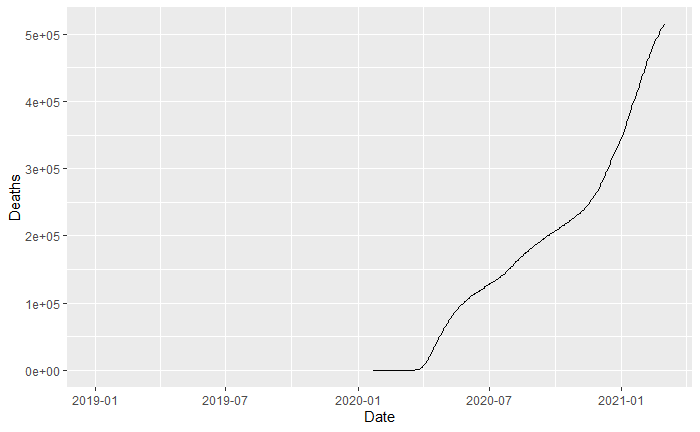
**TSA and COVID 19 Comparison**





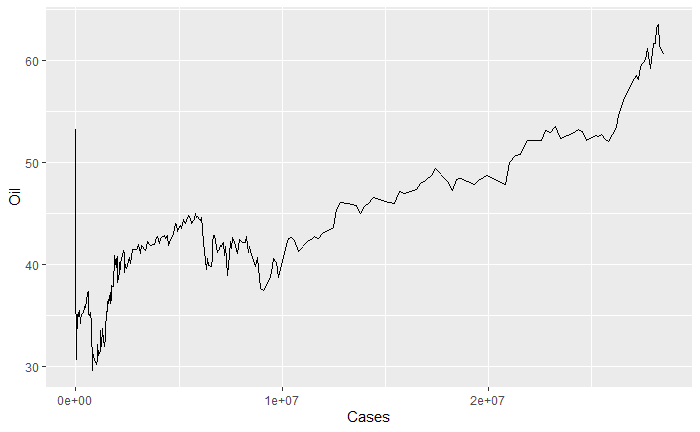
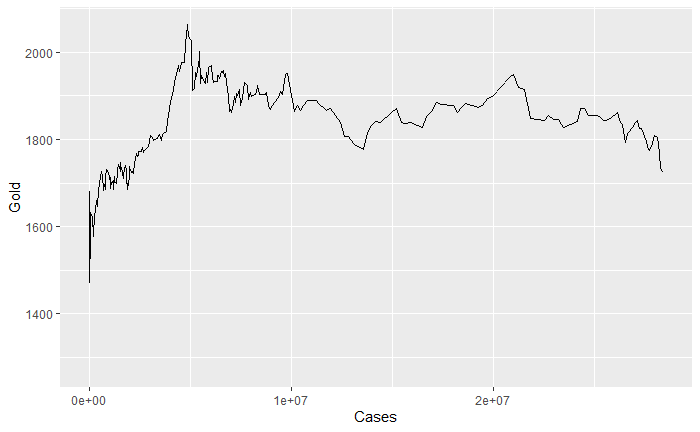
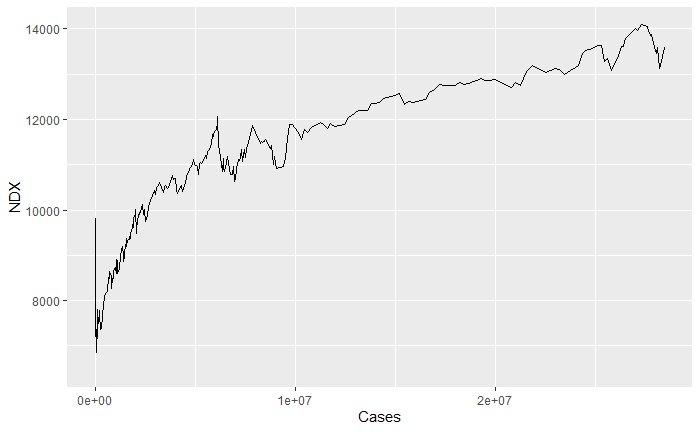
Based on the above initial data, it is clear to see that the amount of people passing through TSA at the airport significantly decreased as a result of COVID 19. The amount of people passing through TSA immediately began to pick up as travel restrictions began to become lifted, however numbers are still well below pre pandemic levels. In addition it is clear to see that there is a positive correlation between the total number of cases and the total number of deaths. However, as the Cases rose above approximately 2.5 million, the death rate per 100,000 cases started to decrease significantly, which could be indicative of several things: treatment of the symptoms improved as healthcare workers gained more knowledge and experience with cases, there were more resources mobilized to treat cases, the virus had potentially mutated to a less deadly strain, the susceptible/exposed population is more resilient (i.e. older/at risk population that was going to be exposed is already dead and only younger/less risky population is still being exposed).

**Covid-19 Deaths, NDX, Gold and Oil Prices over Time**

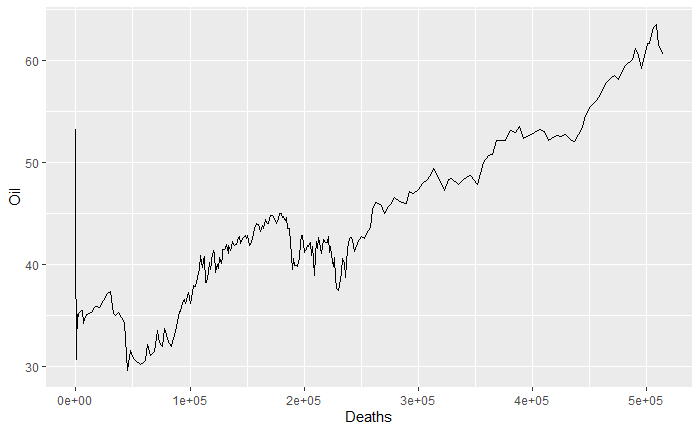
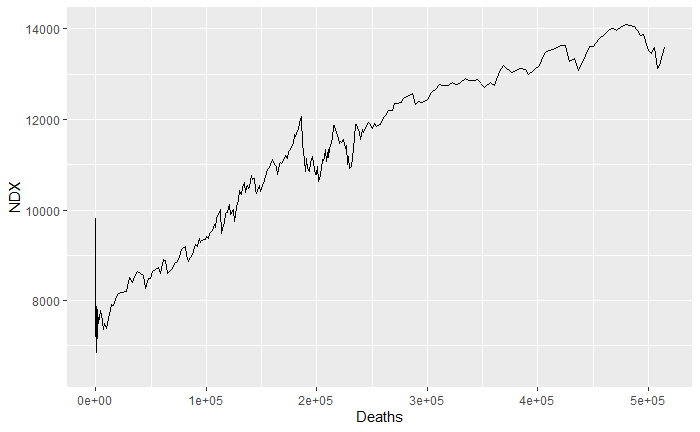


There is a clear dip in the Nasdaq (NDX) and Oil prices at the start of the Covid-19 pandemic. There is also a small, brief dip in Gold price as well, but it rapidly recovers. At this point, the NDX and Oil prices have recovered as well. All three prices seem to have maintained their original trend. This could lead to the hypothesis that all three prices are not positively correlated to Covid-19 deaths, at least not directly.

**NDX, Gold and Oil Prices vs. Covid-19 Cases**



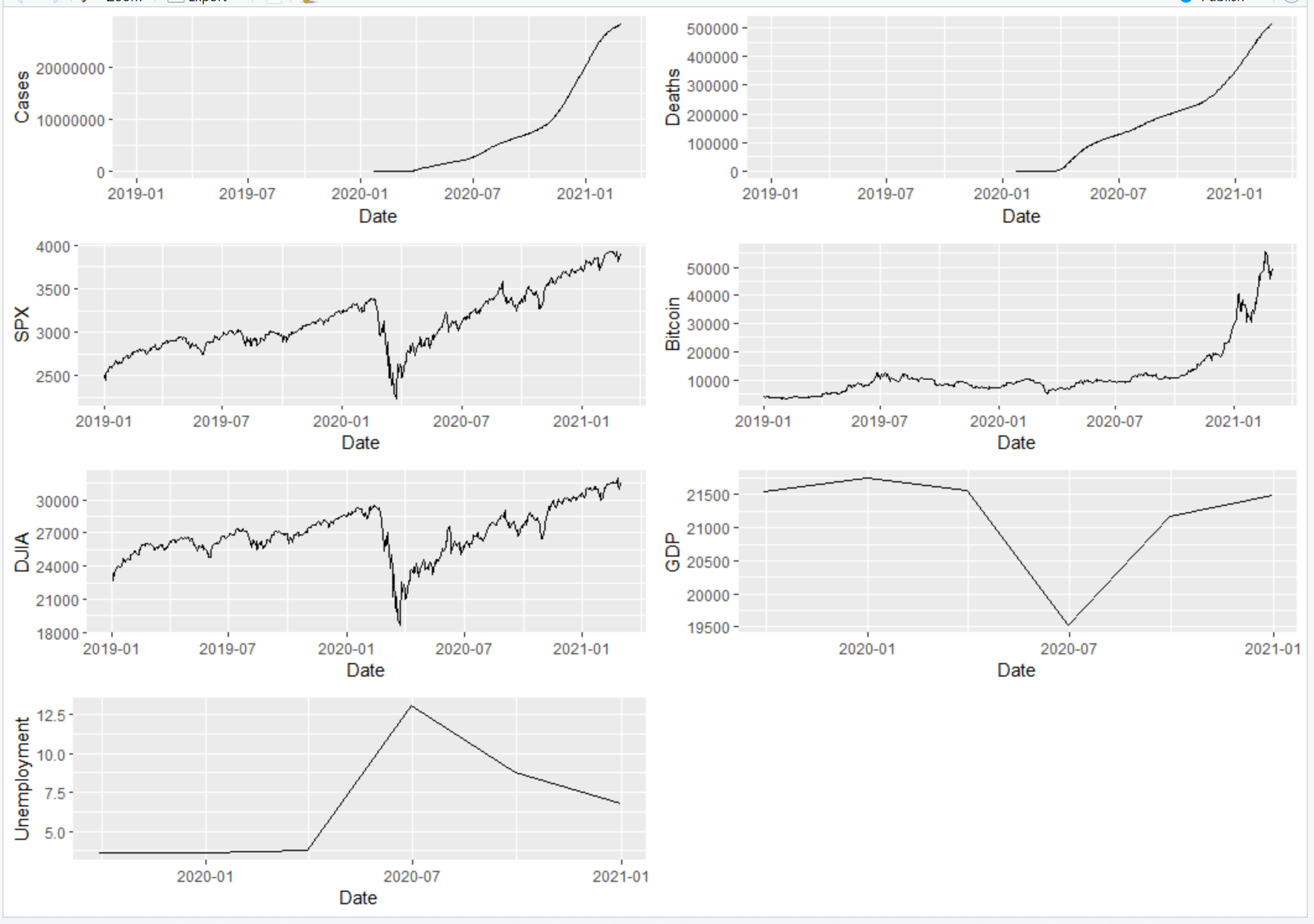
**NDX, Gold and Oil Prices vs. Covid-19 Deaths**



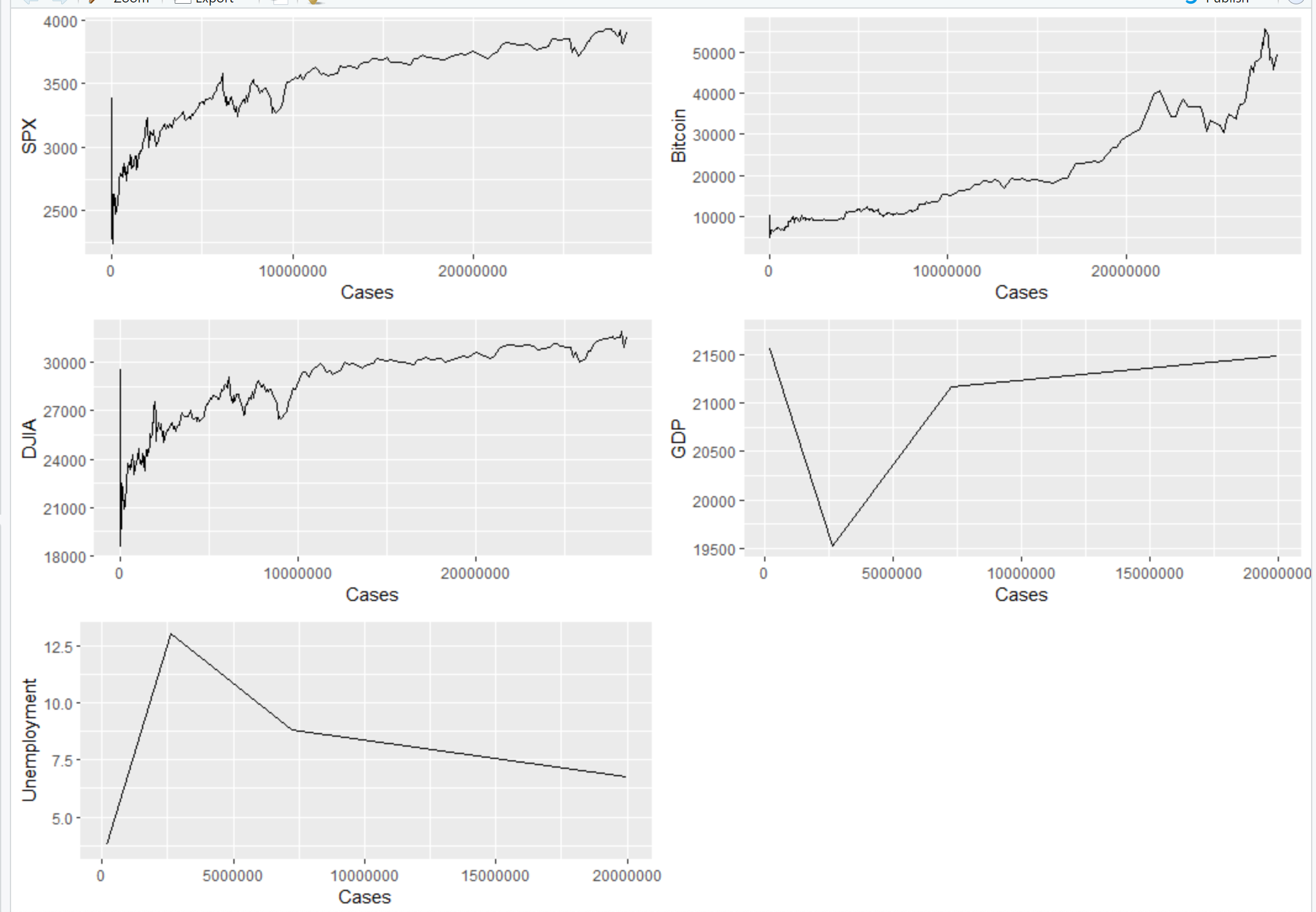
It can be seen from plots of prices vs. Cases or Death do not show any real correlation. One could make the assumption that the prices of NDX and Oil are positively correlated with both Cases and Deaths but that is more like just due to price increase over time, due to other factors.

There may be another factor at play, e.g. the initial news of the pandemic caused a scare that settled over time. Perhaps other factors such as macroeconomic indicators or the world’s rapid improvement in virtualization provided stability to the market.

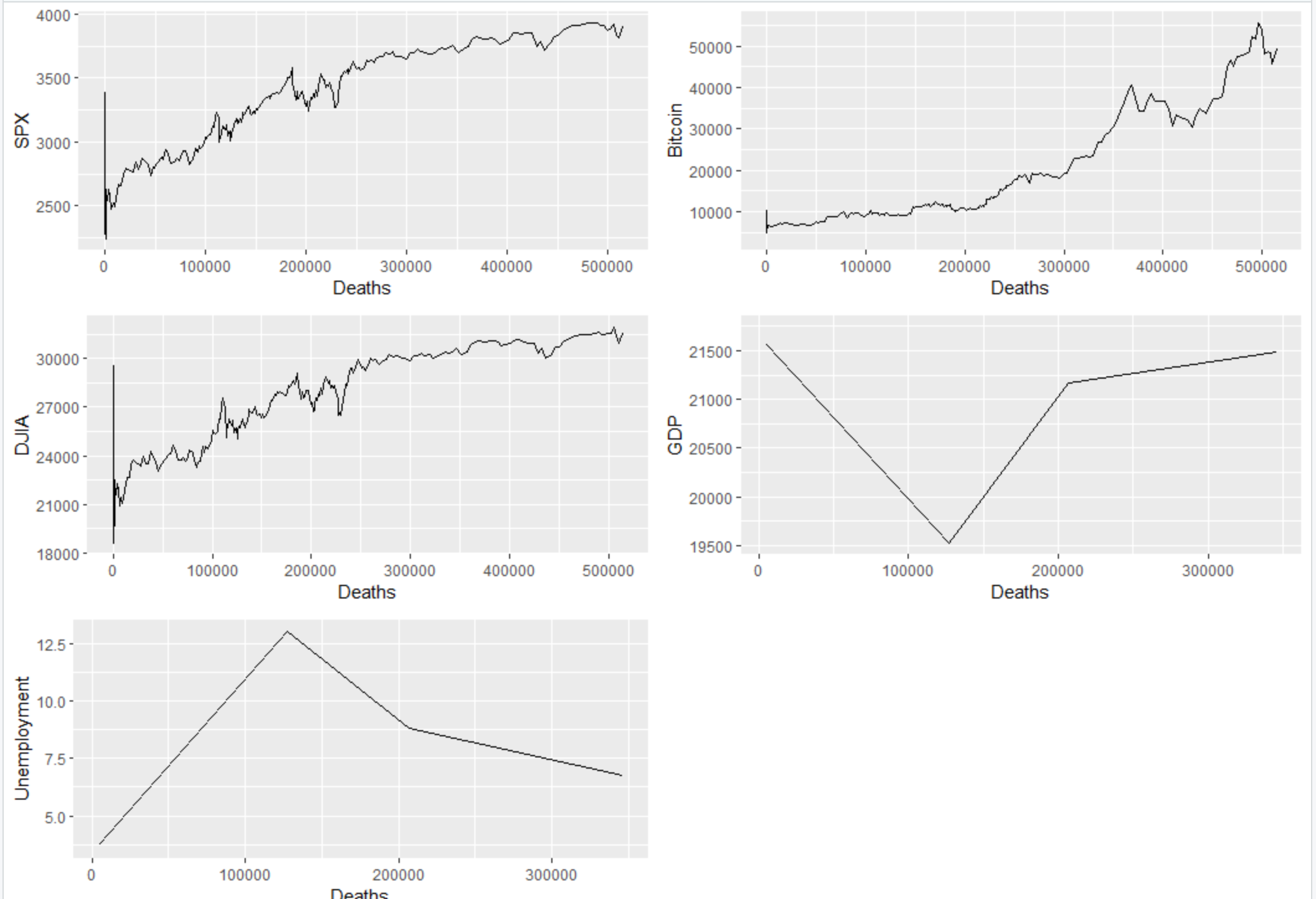
**SPX, DOW, Bitcoin, GDP and Unemployment**



From the plots above we can see that the SPX, Bitcoin, DJIA, GDP and Unemployment dropped significantly and then later rebounding in early 2021.



From the plots above we could see that the BTCUSD pair did not see significant fluctuation in its price during the pandemic, rest of the other markets responded to the case and their price were unstable until the cases hit 10M.



From the plots above we could see that the BTCUSD pair did not see significant fluctuation in its price during the pandemic, rest of the other markets responded to the case and their price were unstable until the death reached 100,000.

**Modeling**

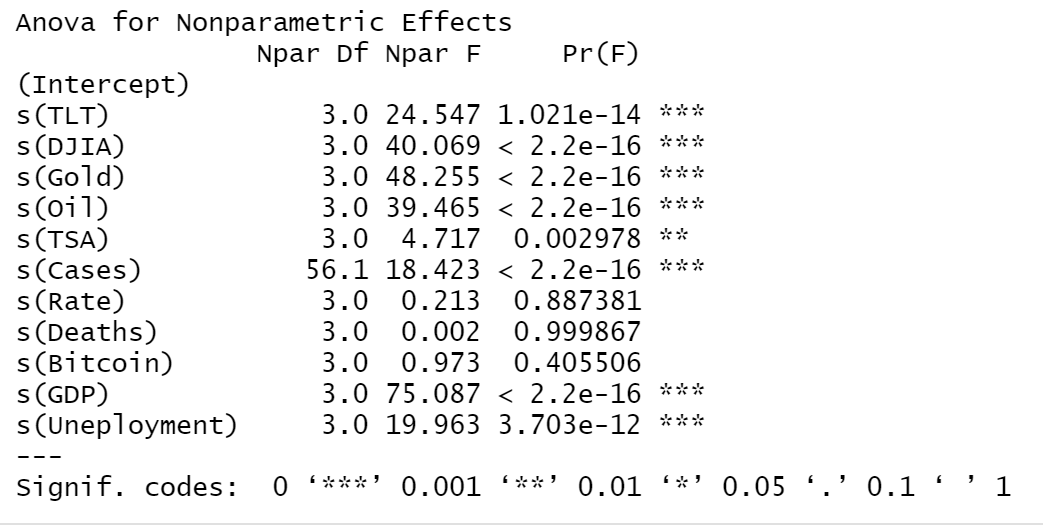
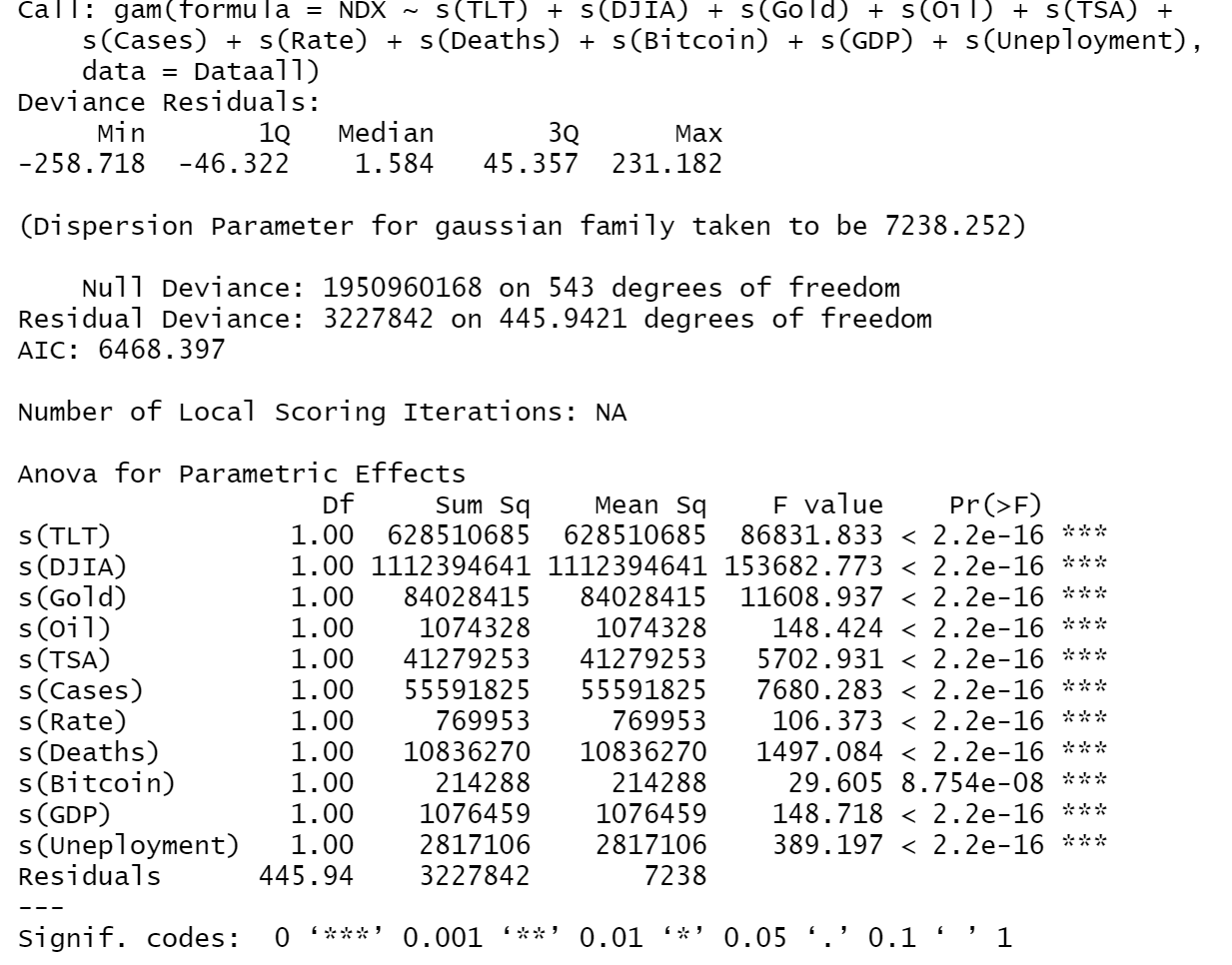
After initially looking at the data and analyzing some of the features and relationships between the data, several different models were developed and explored in order to determine which model was the most appropriate for our data.

**GAM Model**

The first model that was explored was the GAM model. There were four separate models created. The first two attempts used NDX as the output. One of them used all of the data from 2019 to March 1, 2021 while the other used data from only from the start of Covid to March1, 2021. The other two models used DJIA as the output variable and once again split the data into a training and test set. Below the results from all four of the models are displayed.

1. **GAM Model NDX All Data**

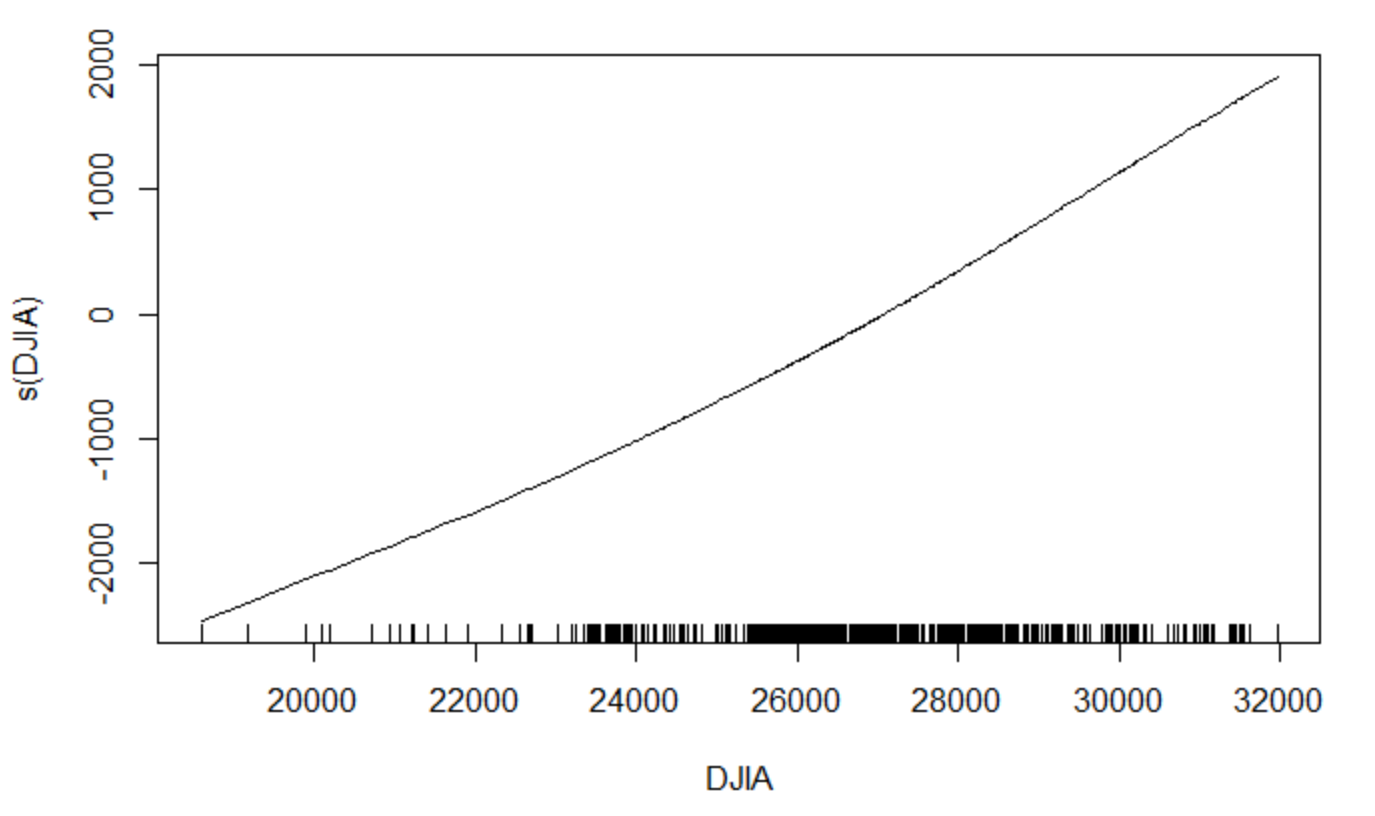
Summary

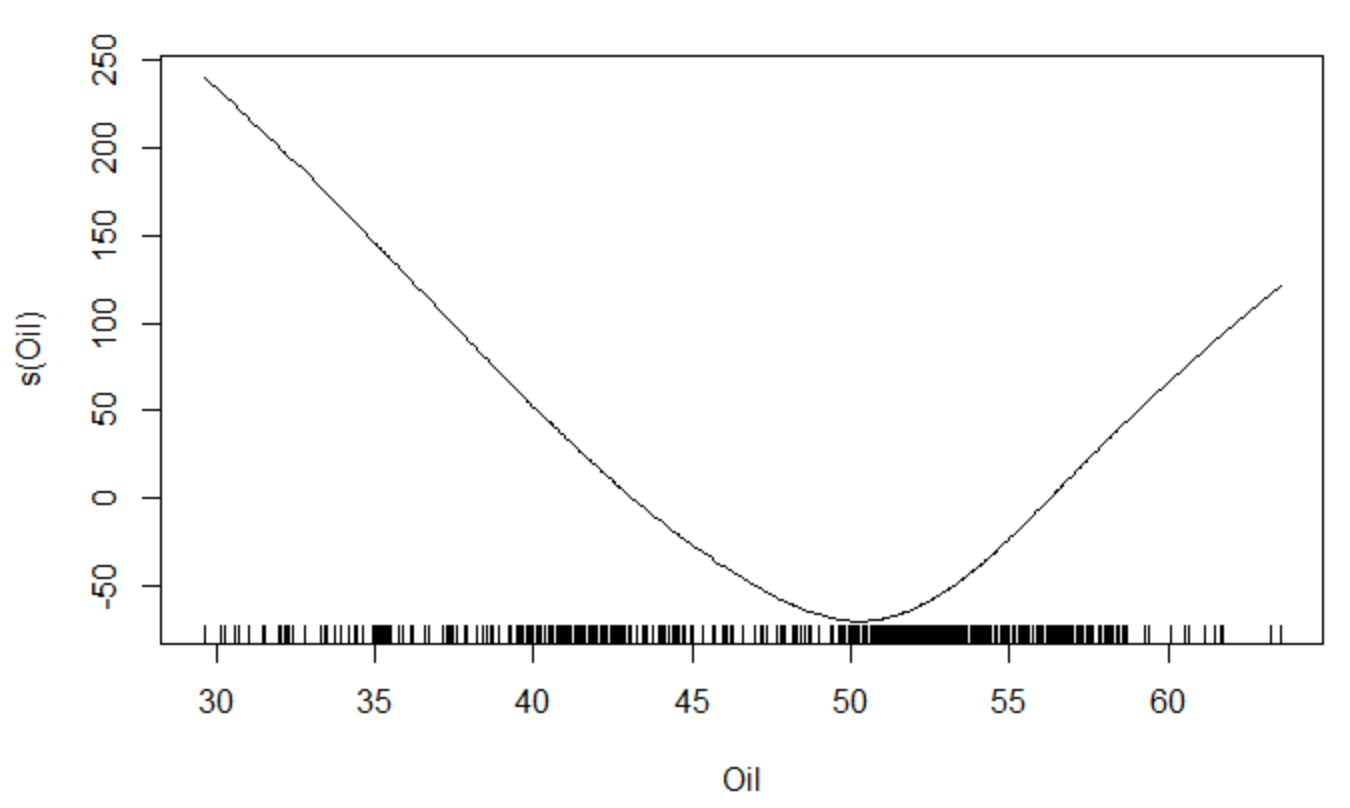
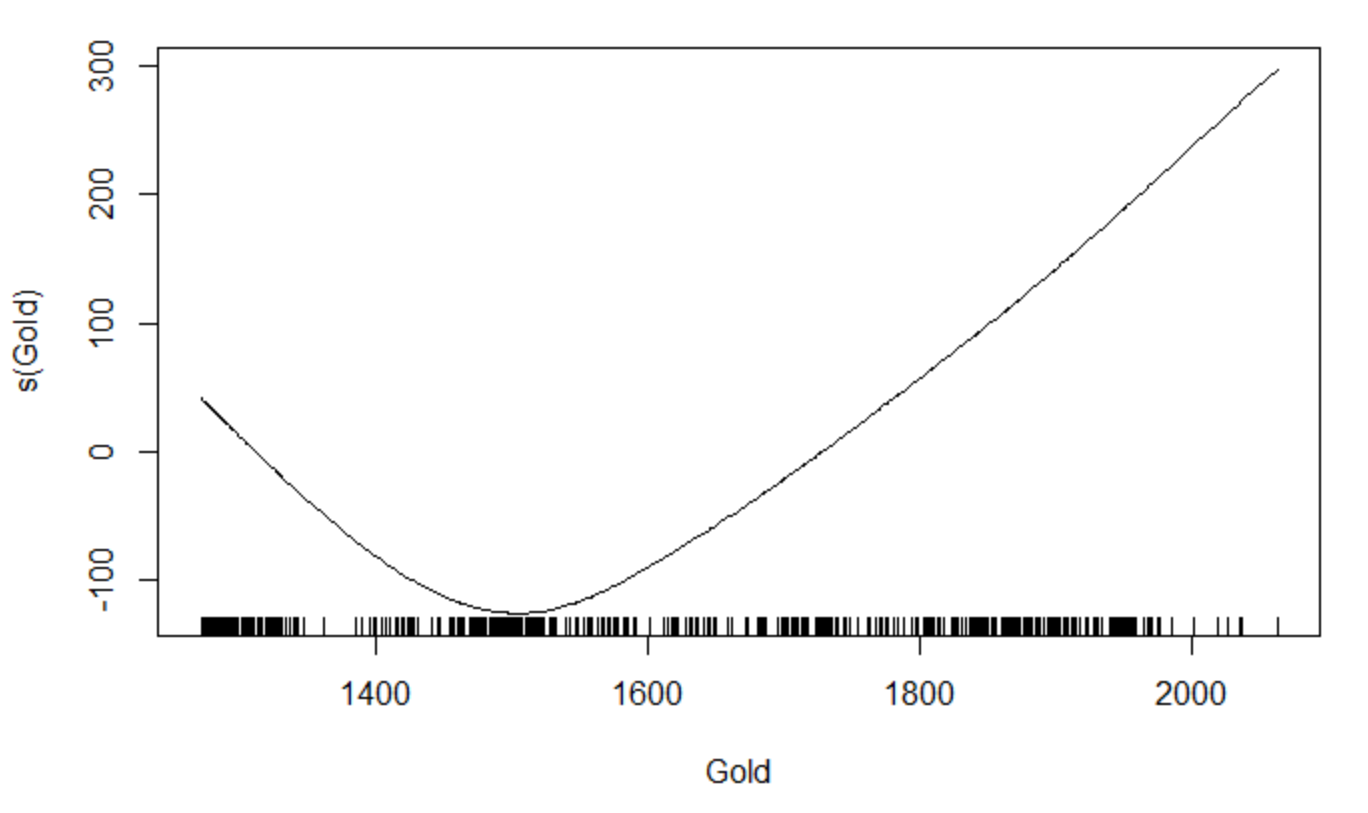


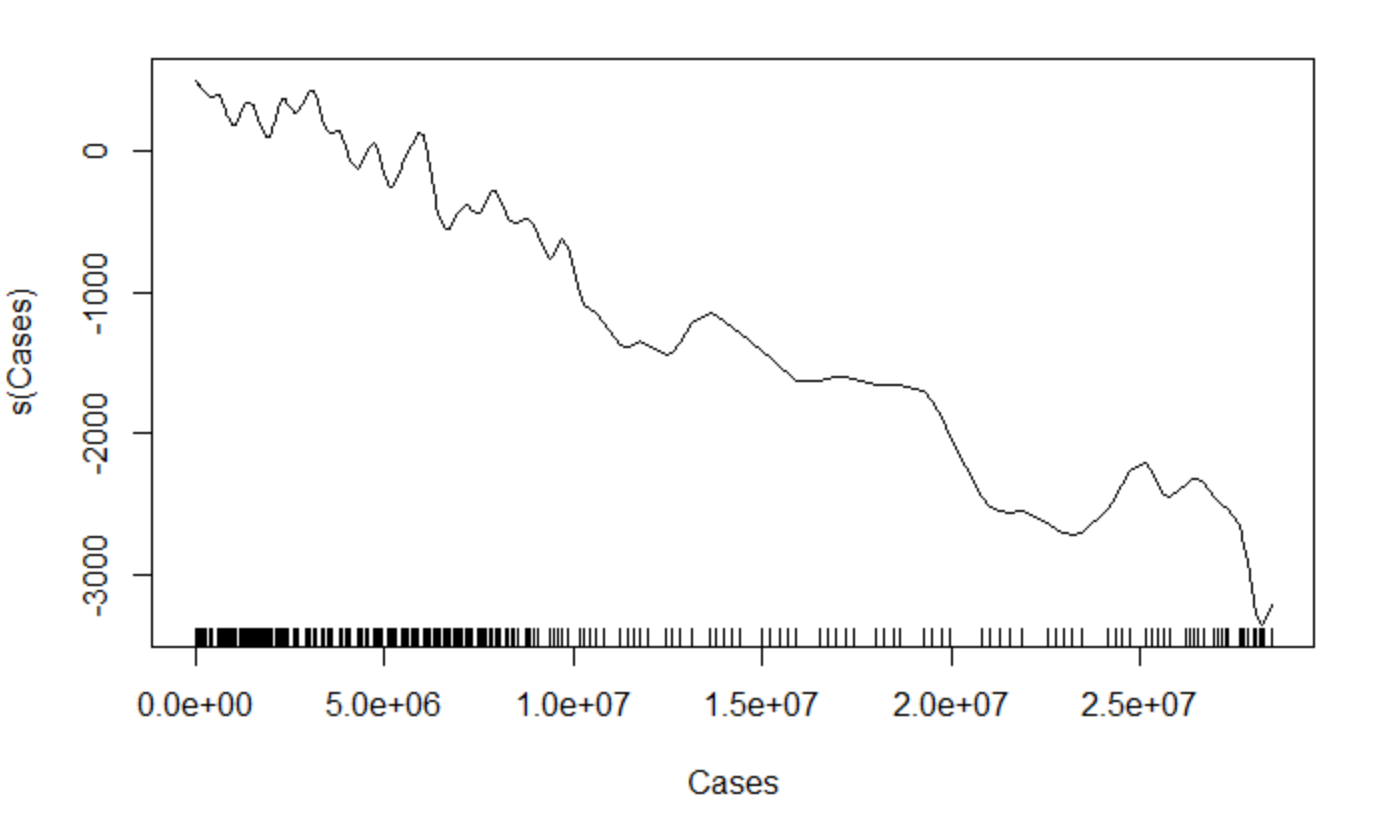
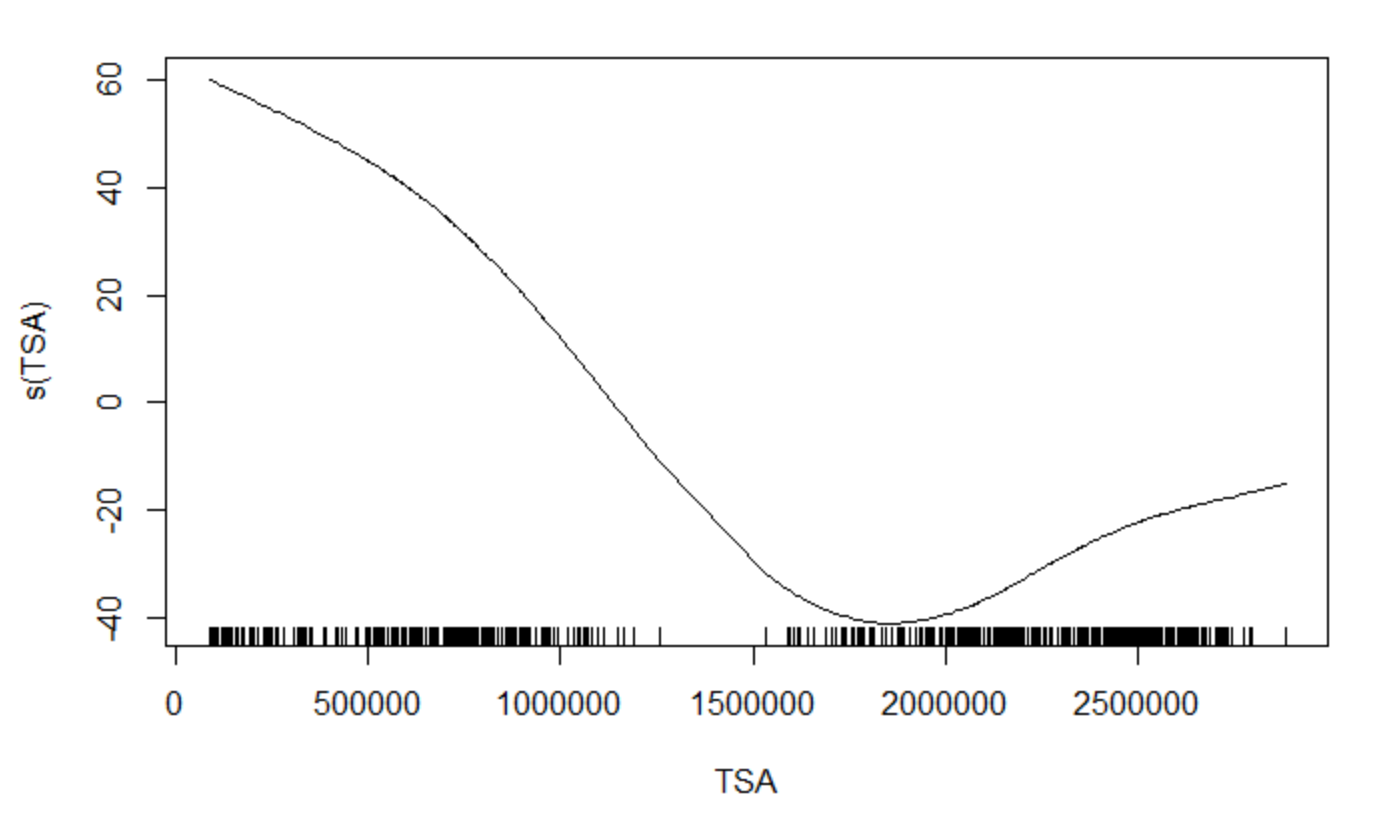
Coefficient

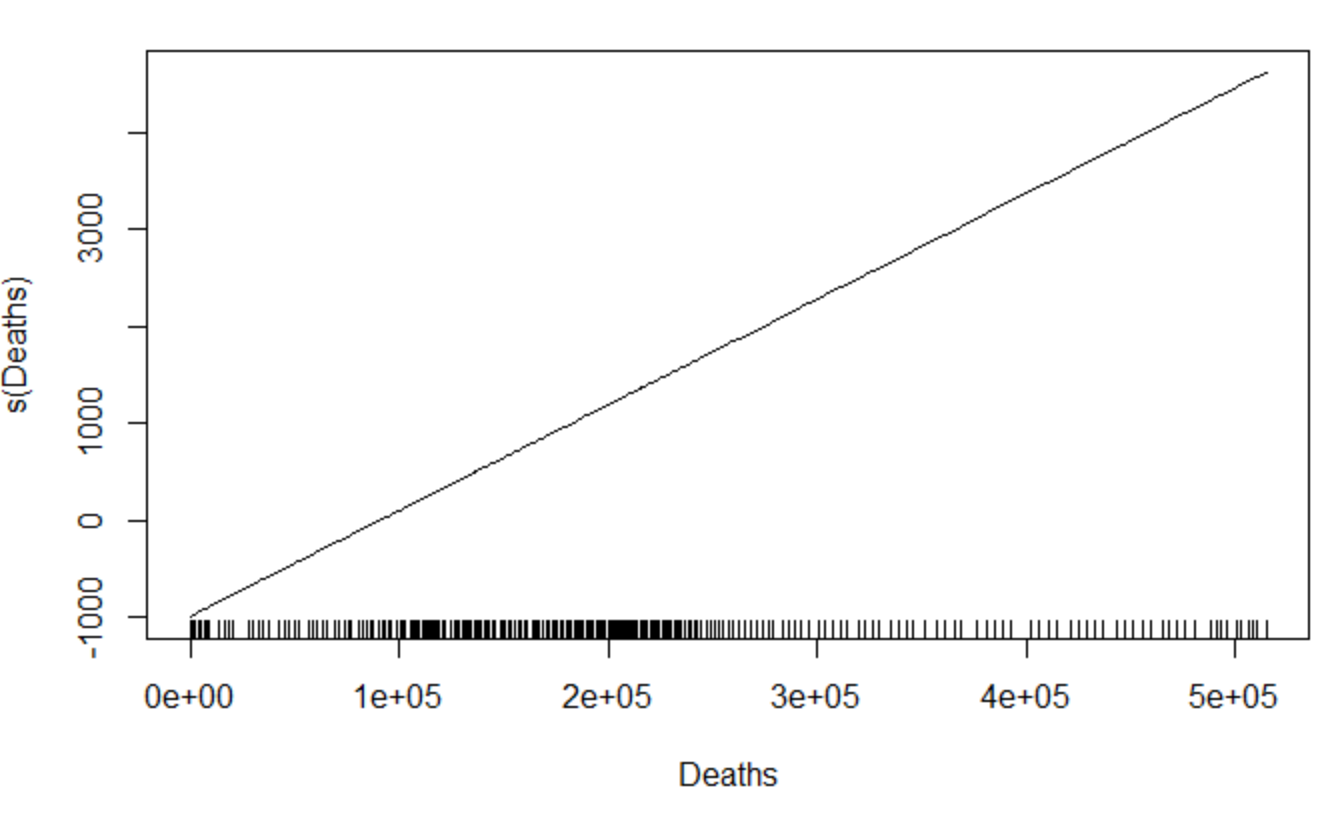
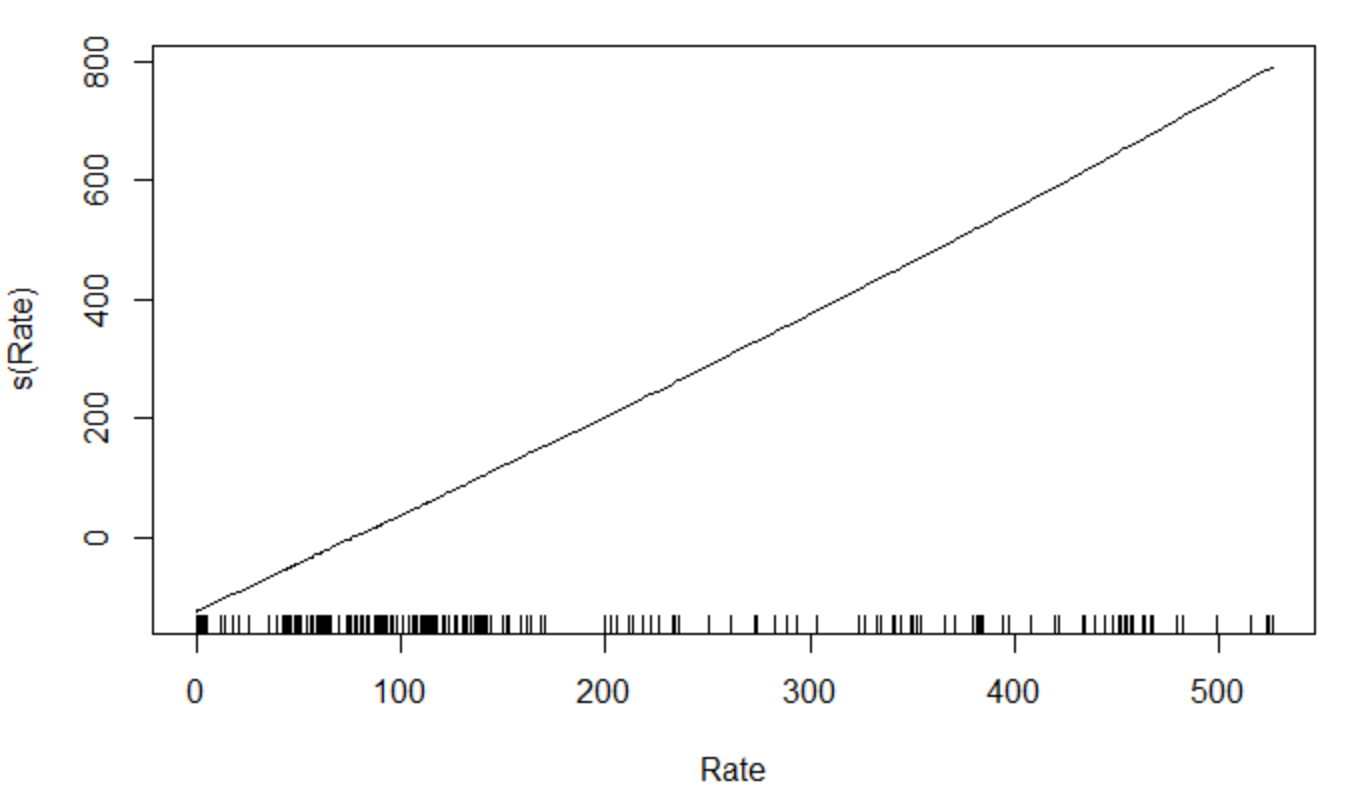


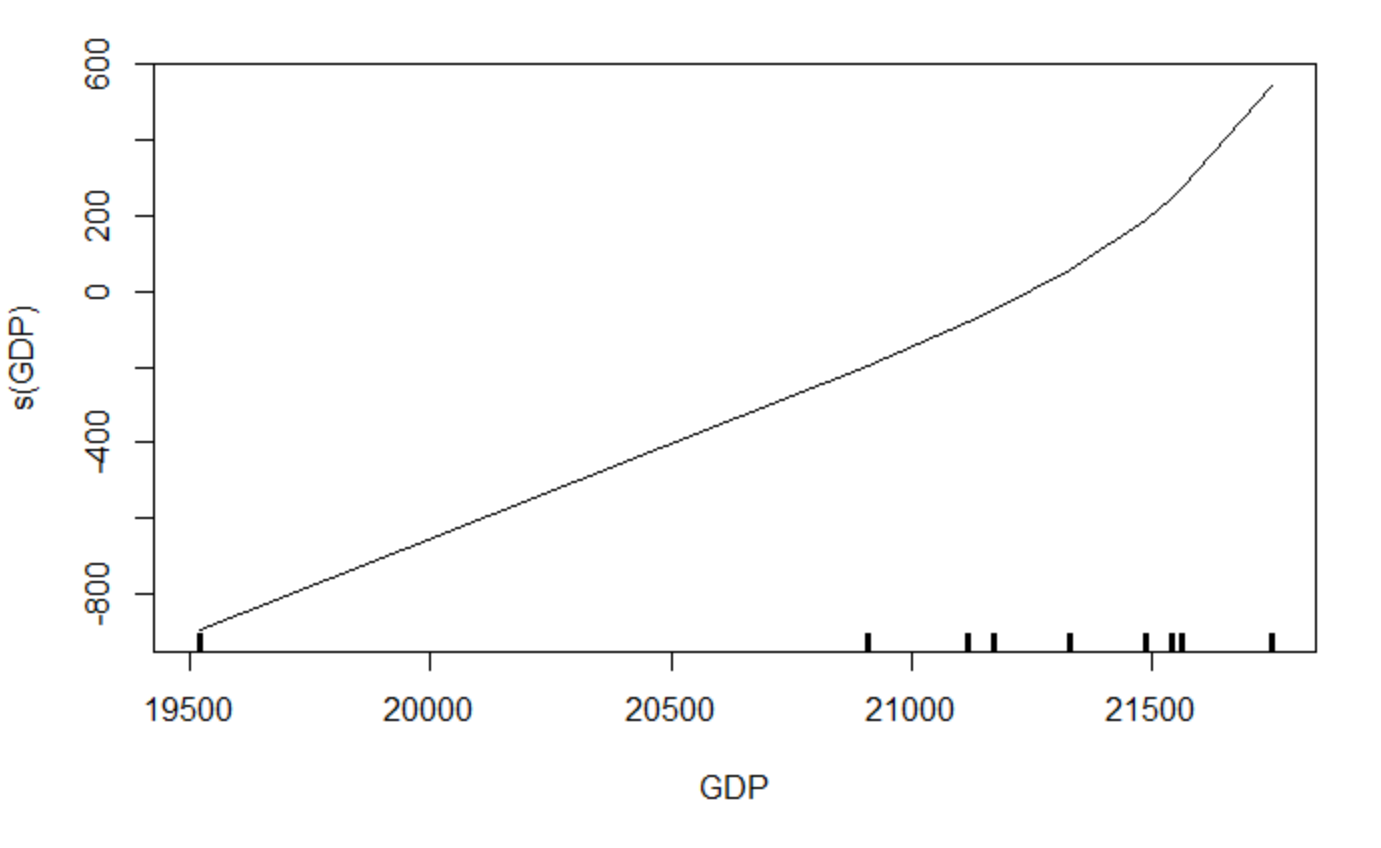
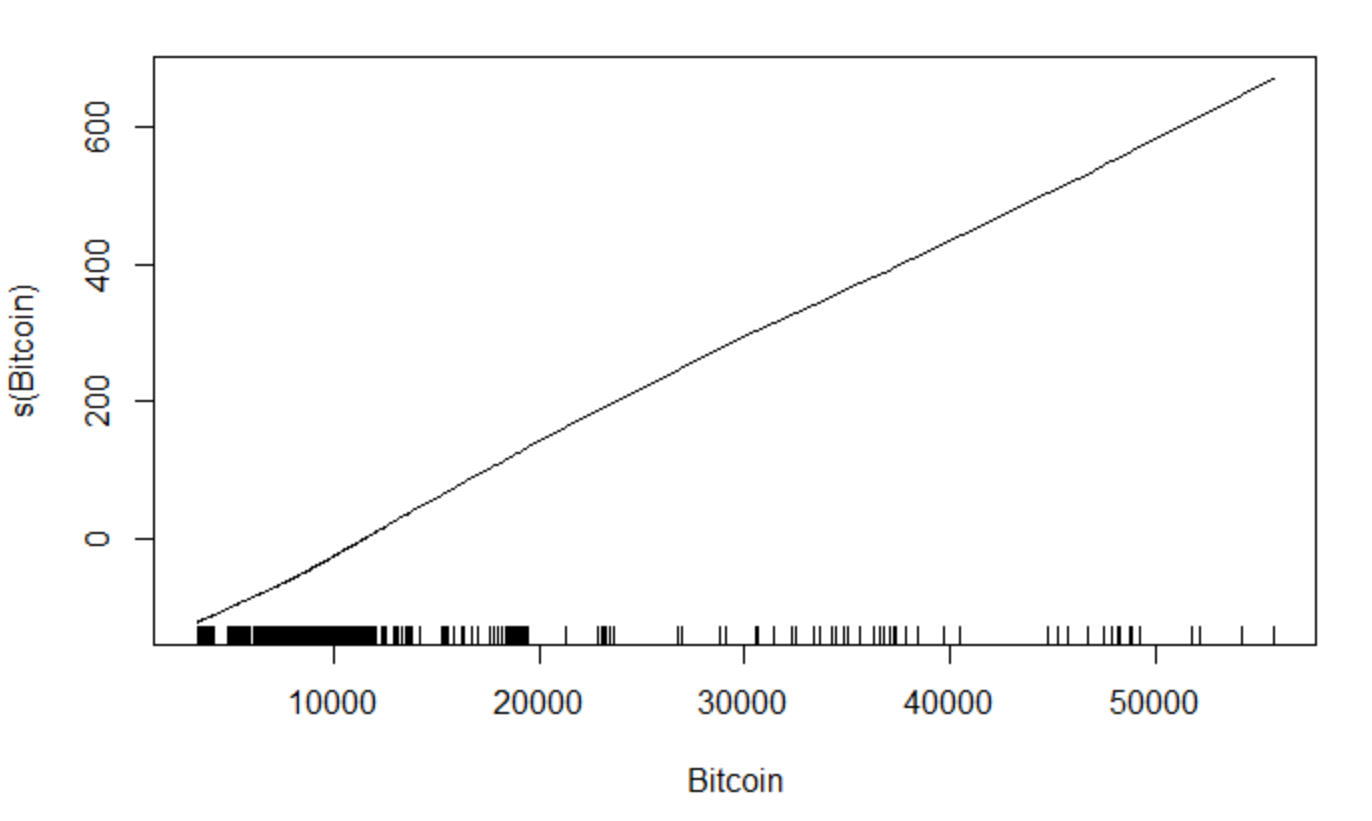
Plots

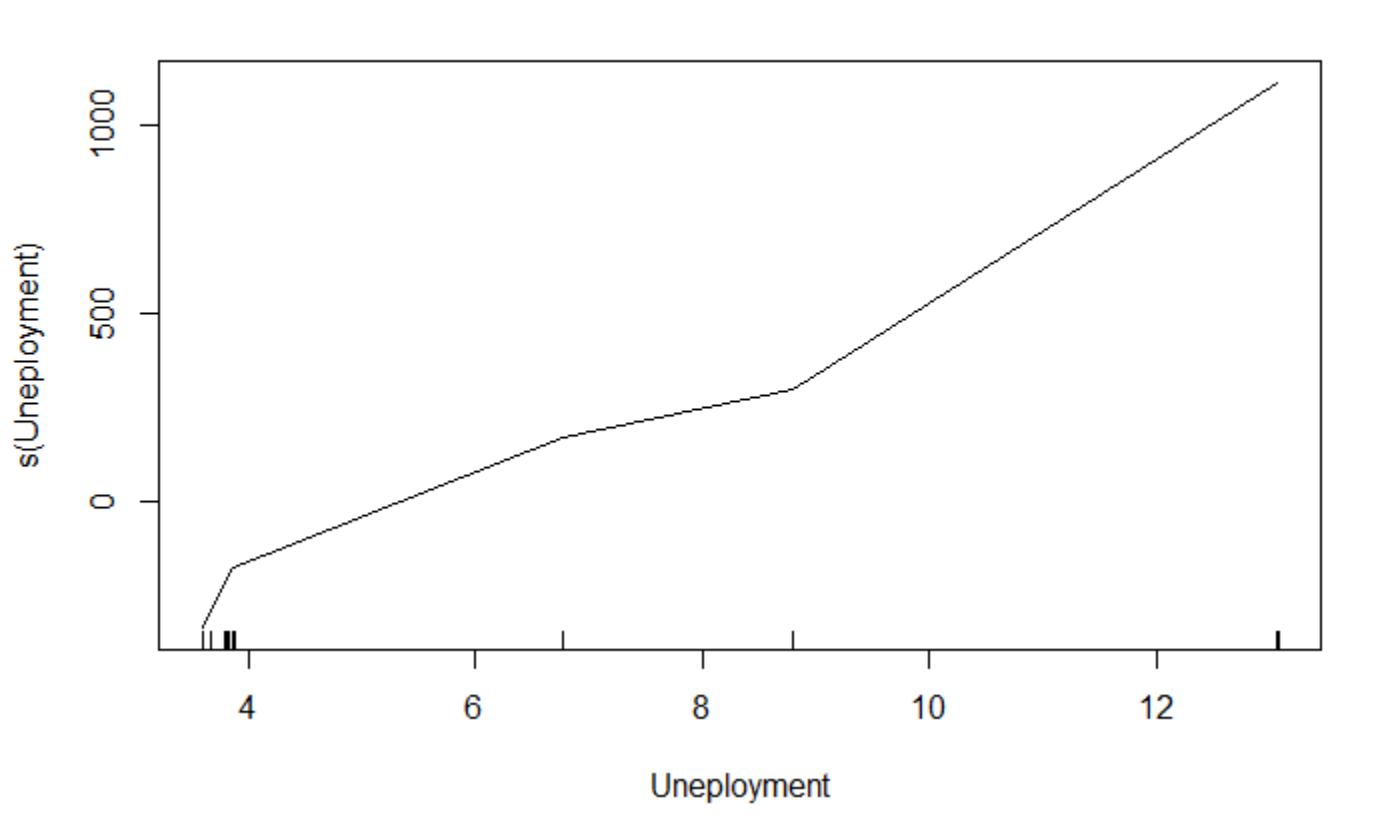






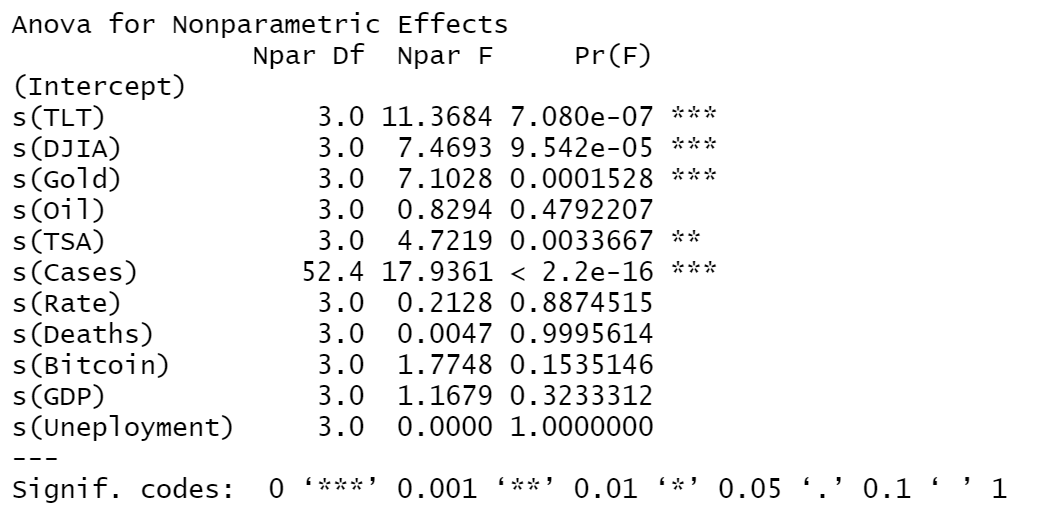
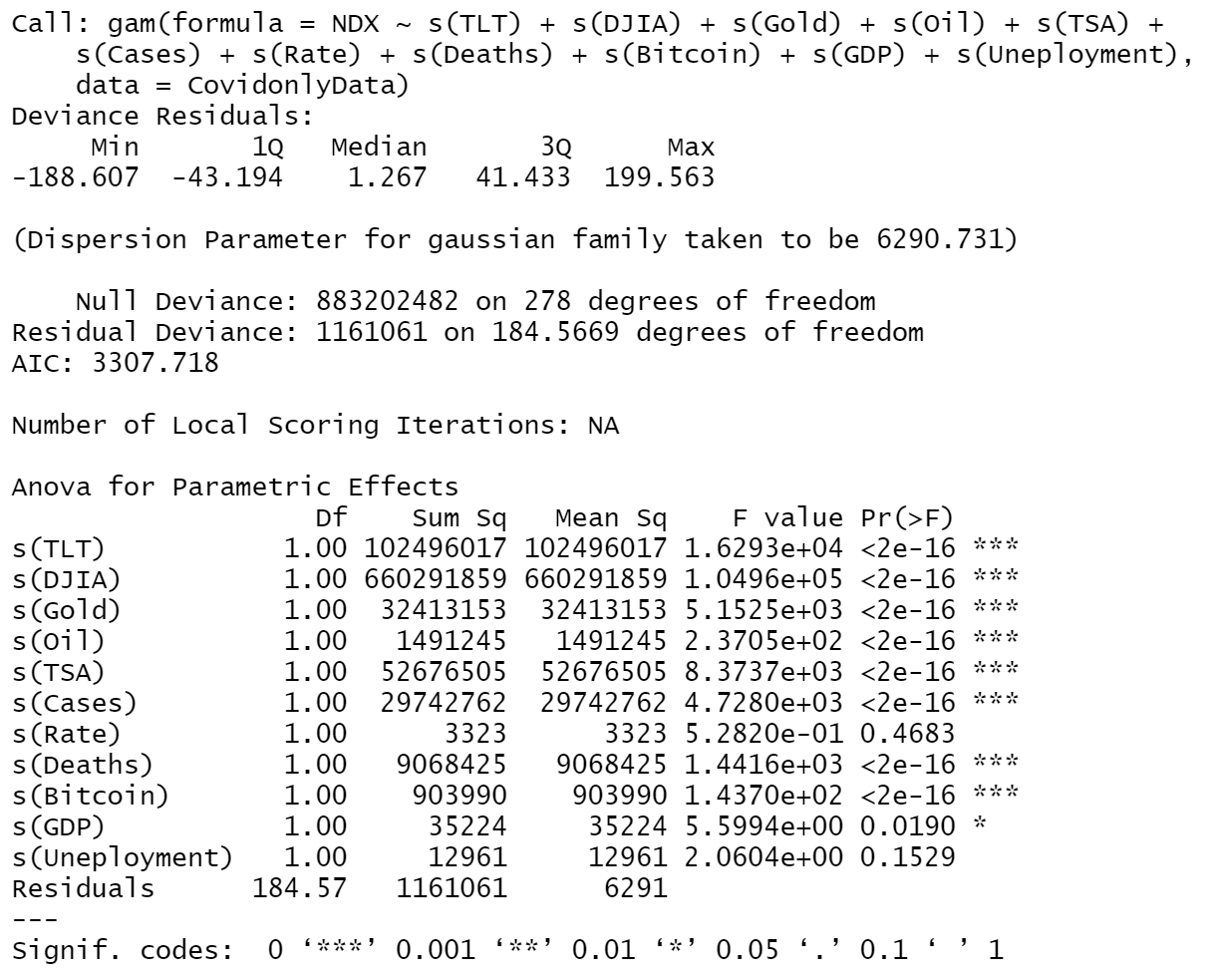




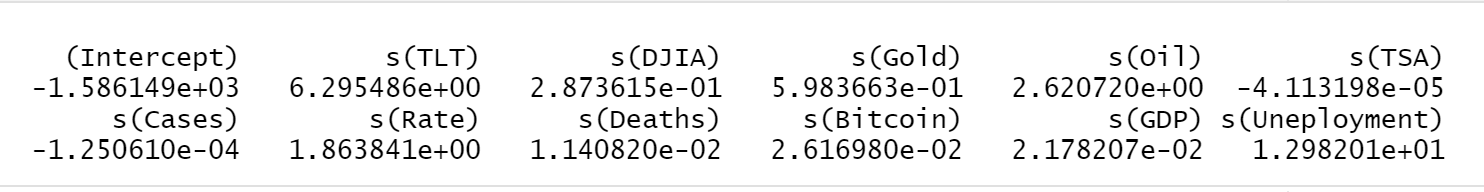


1. **GAM Model NDX Covid Only**

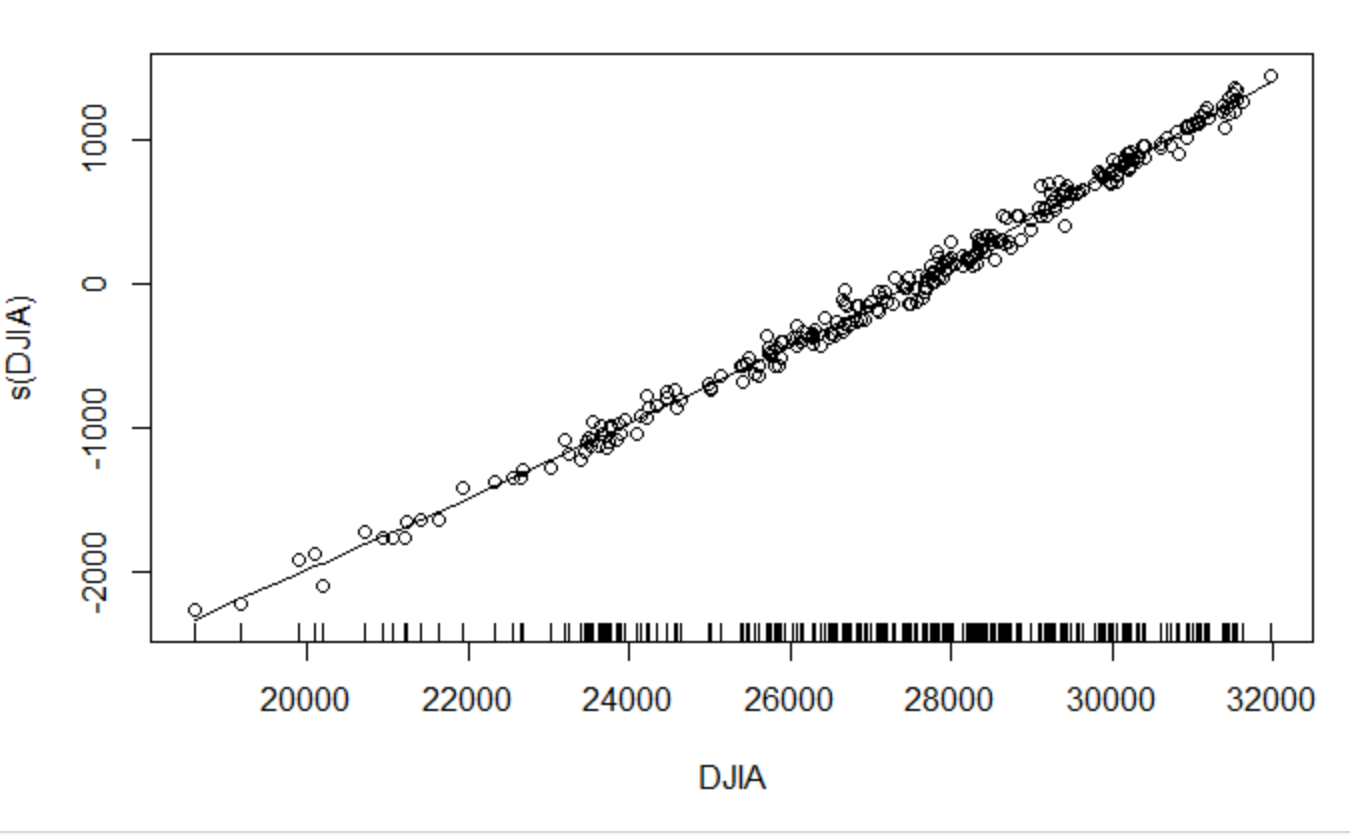
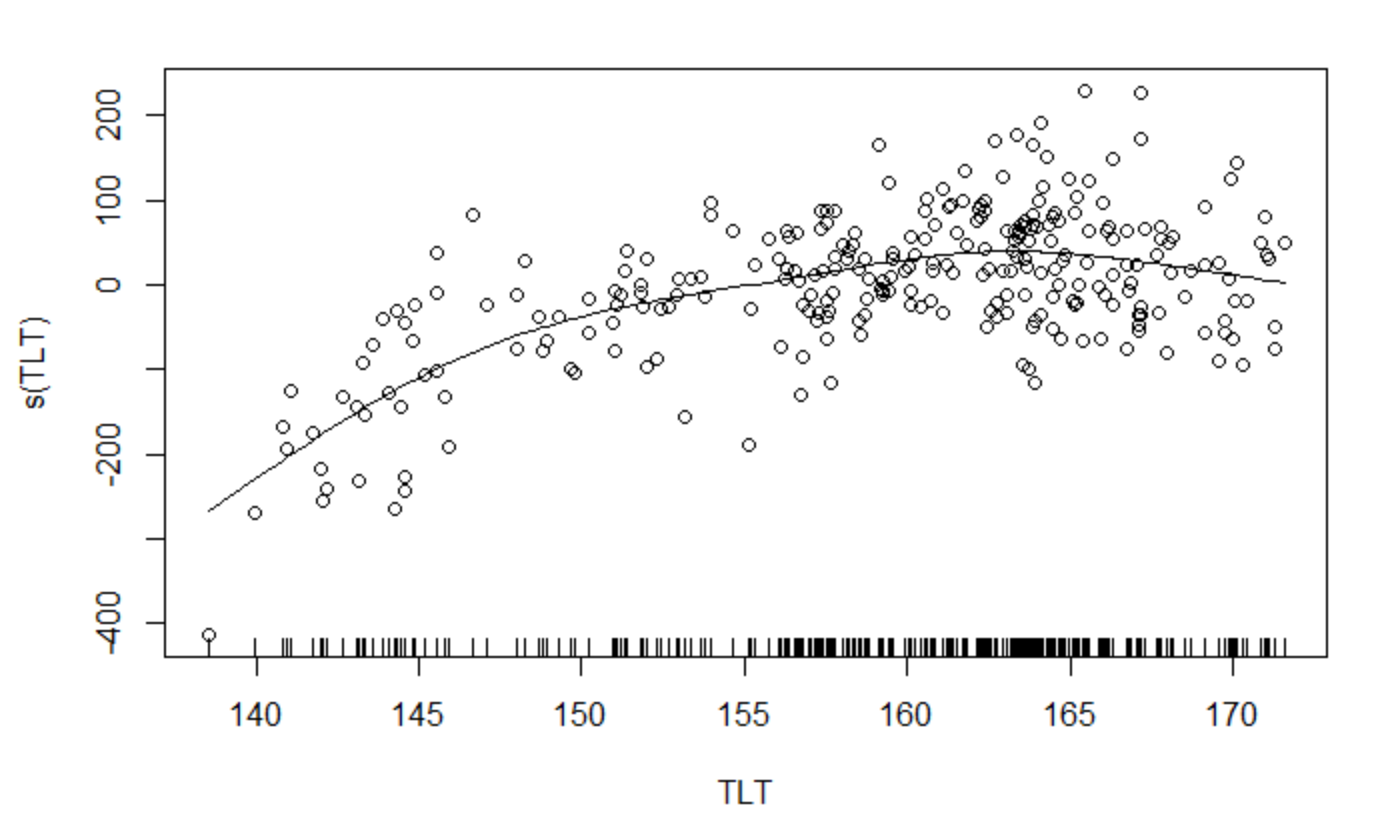
Summary

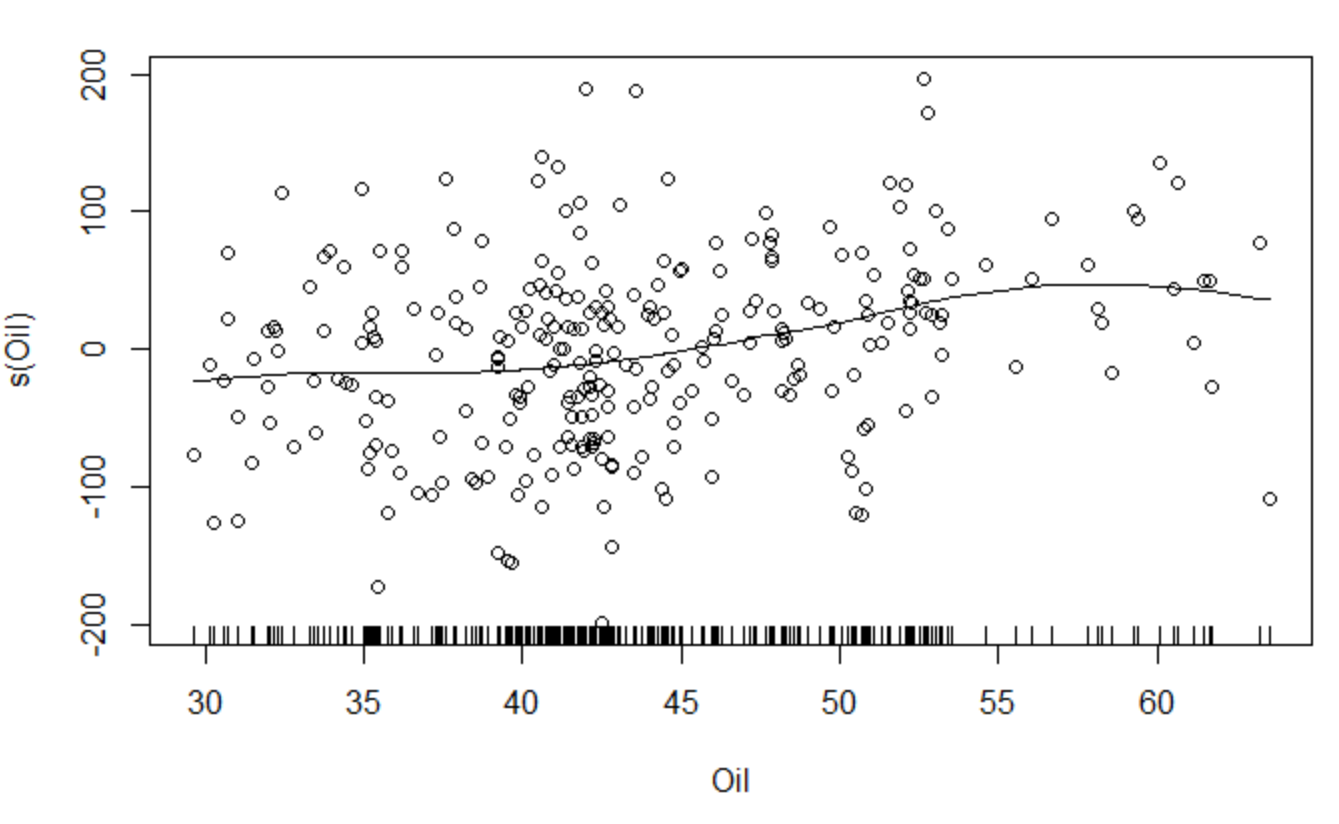
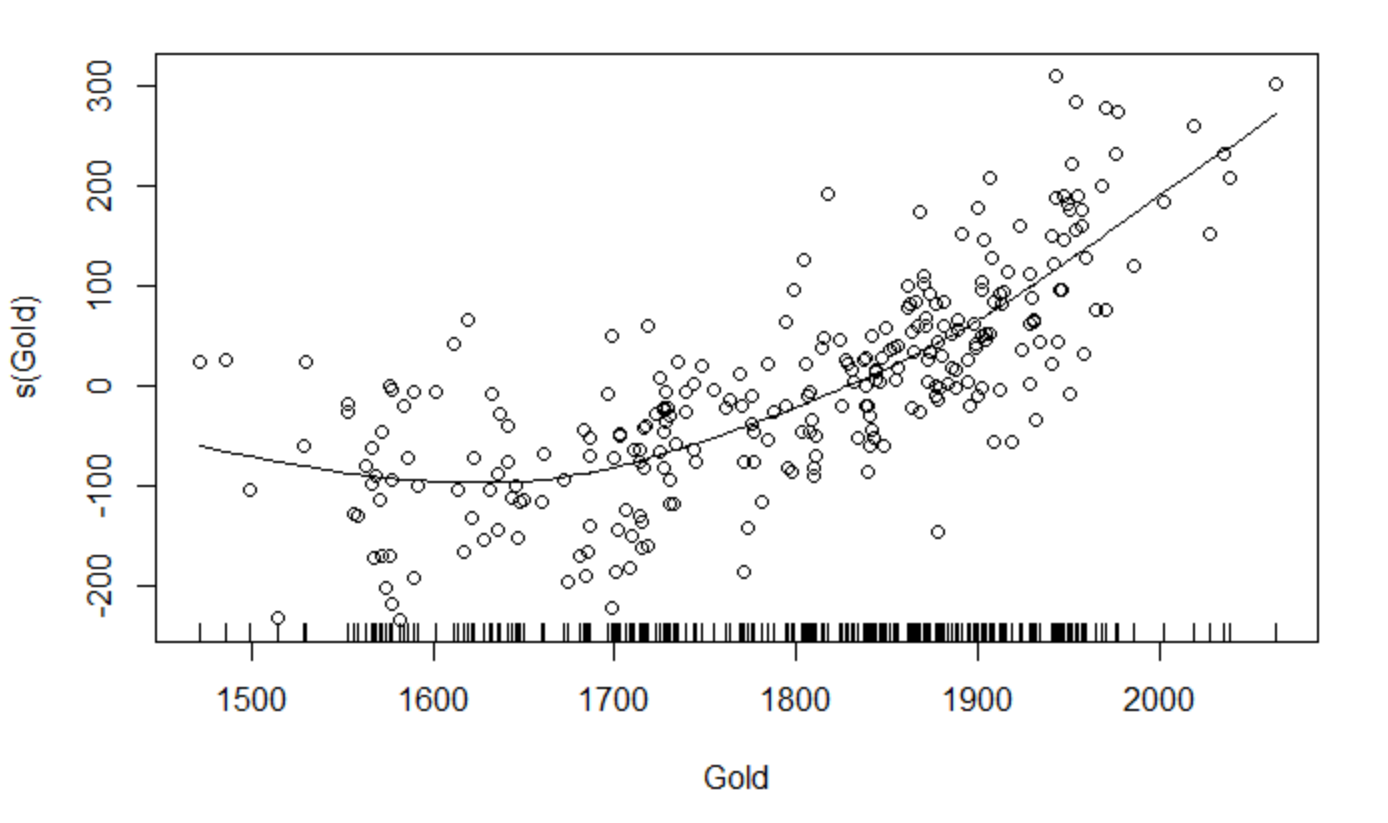


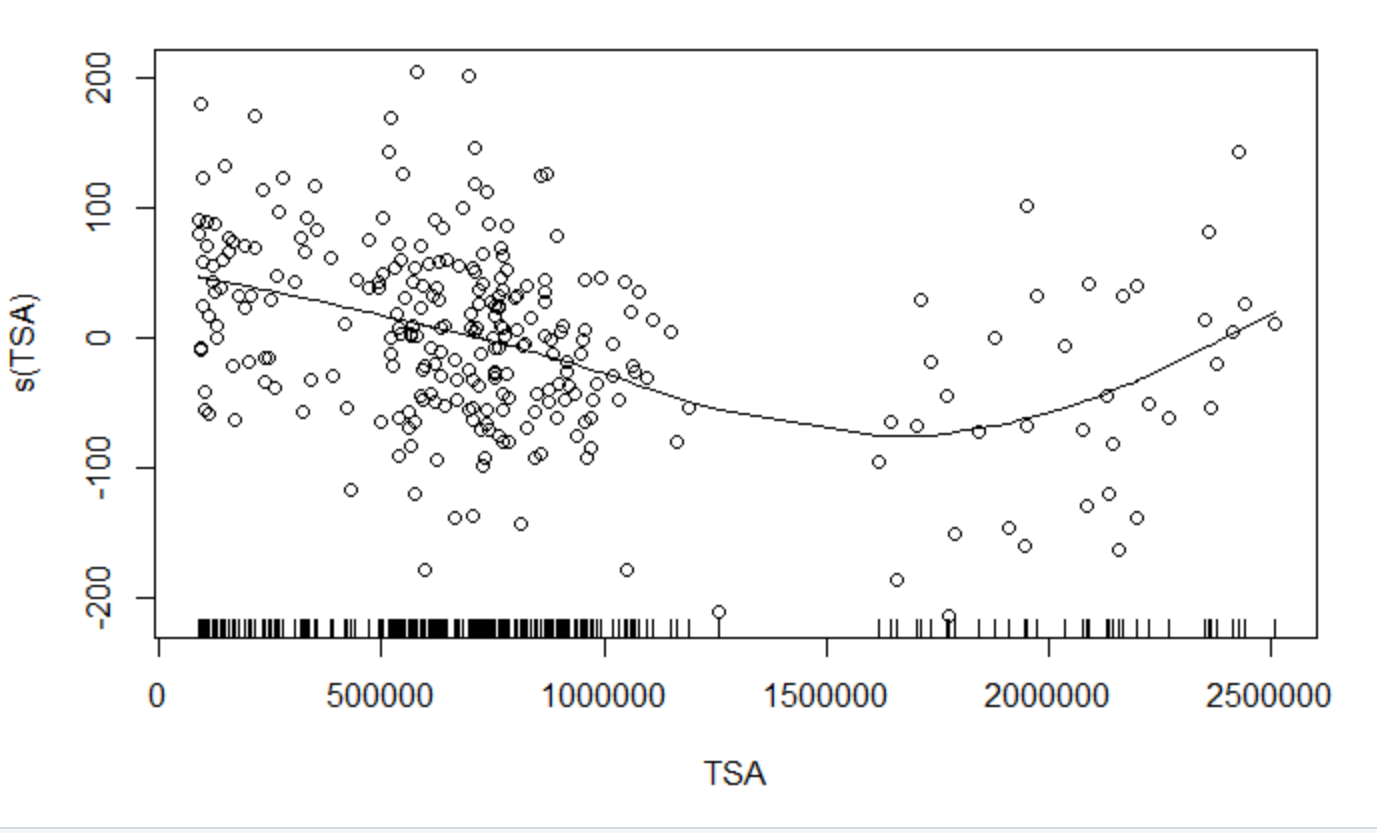
Coefficient

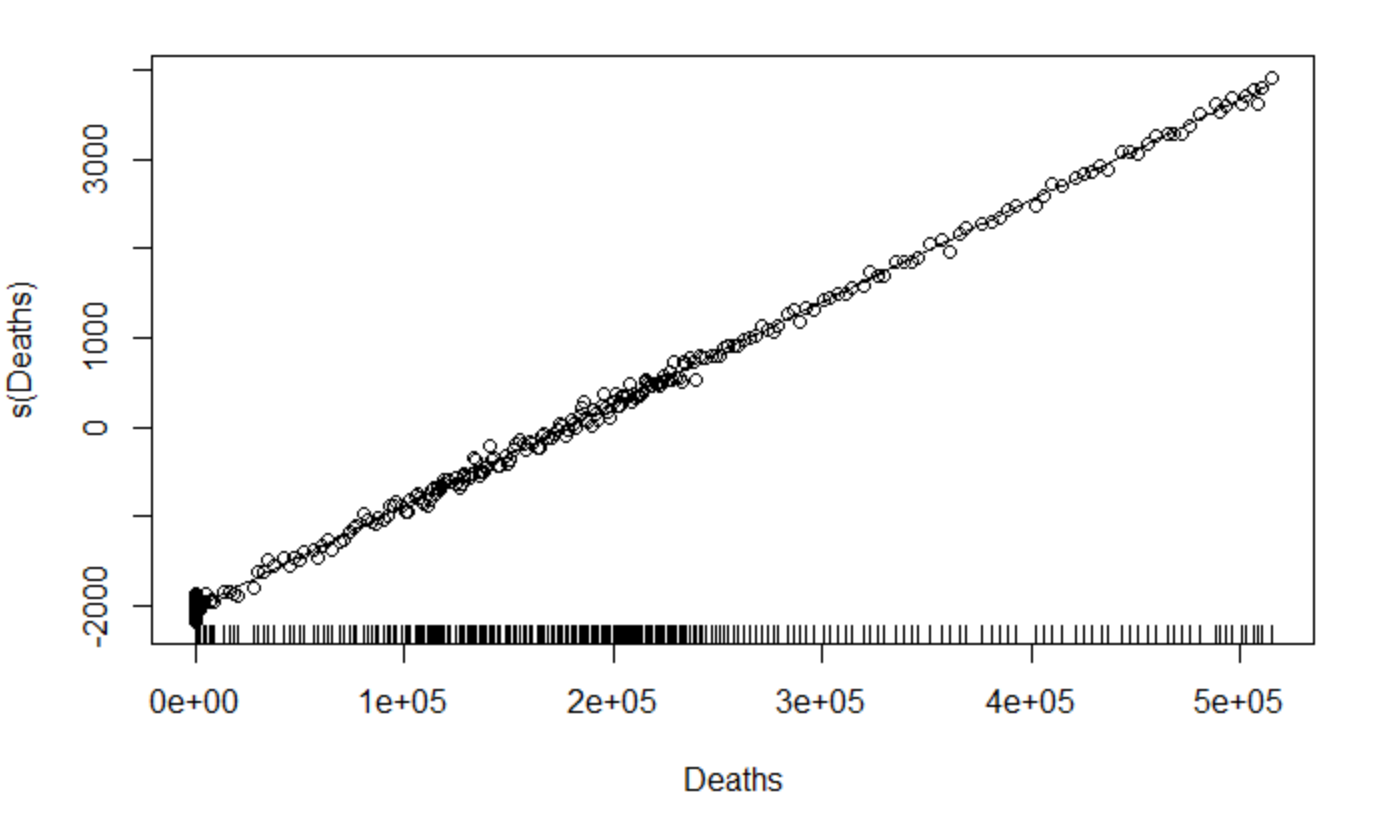
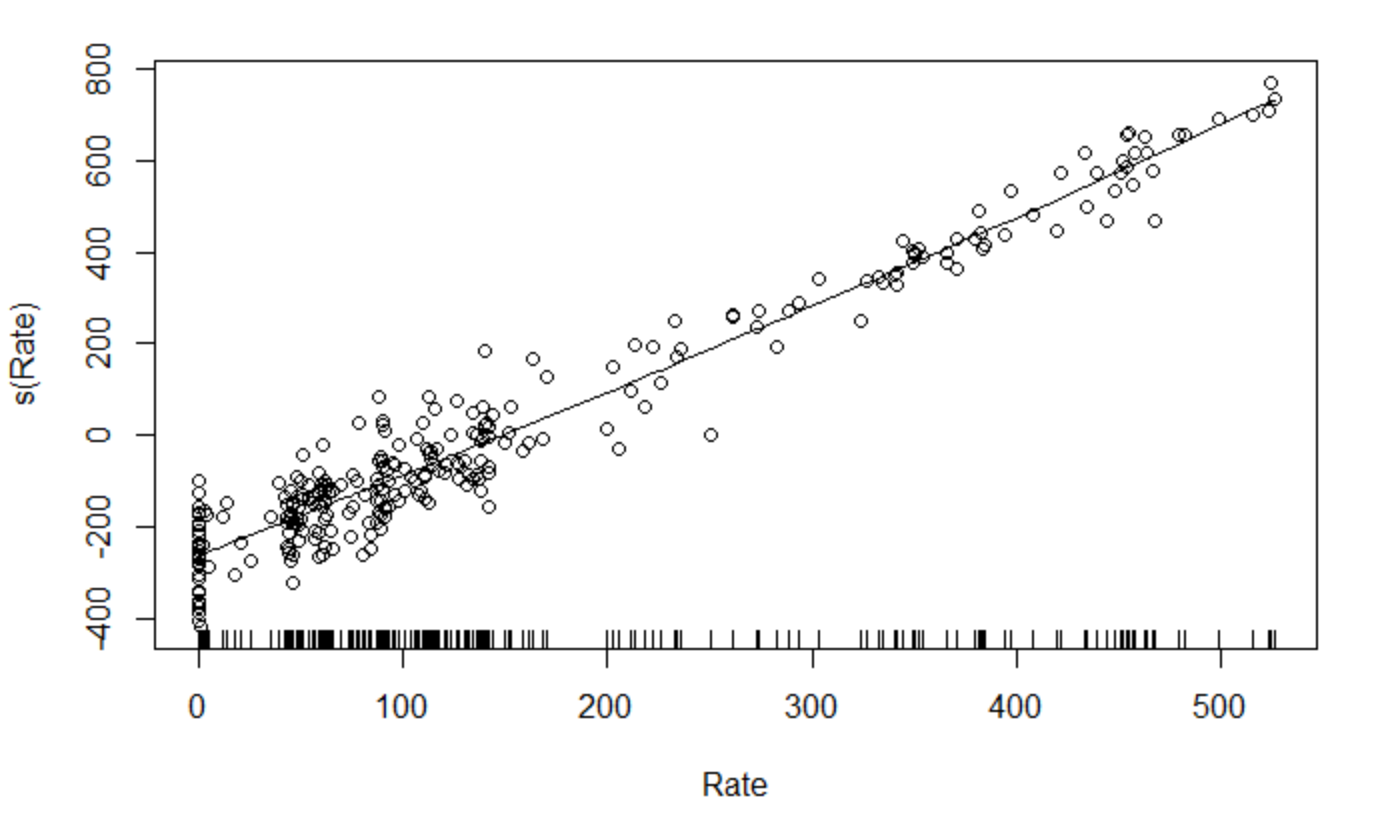


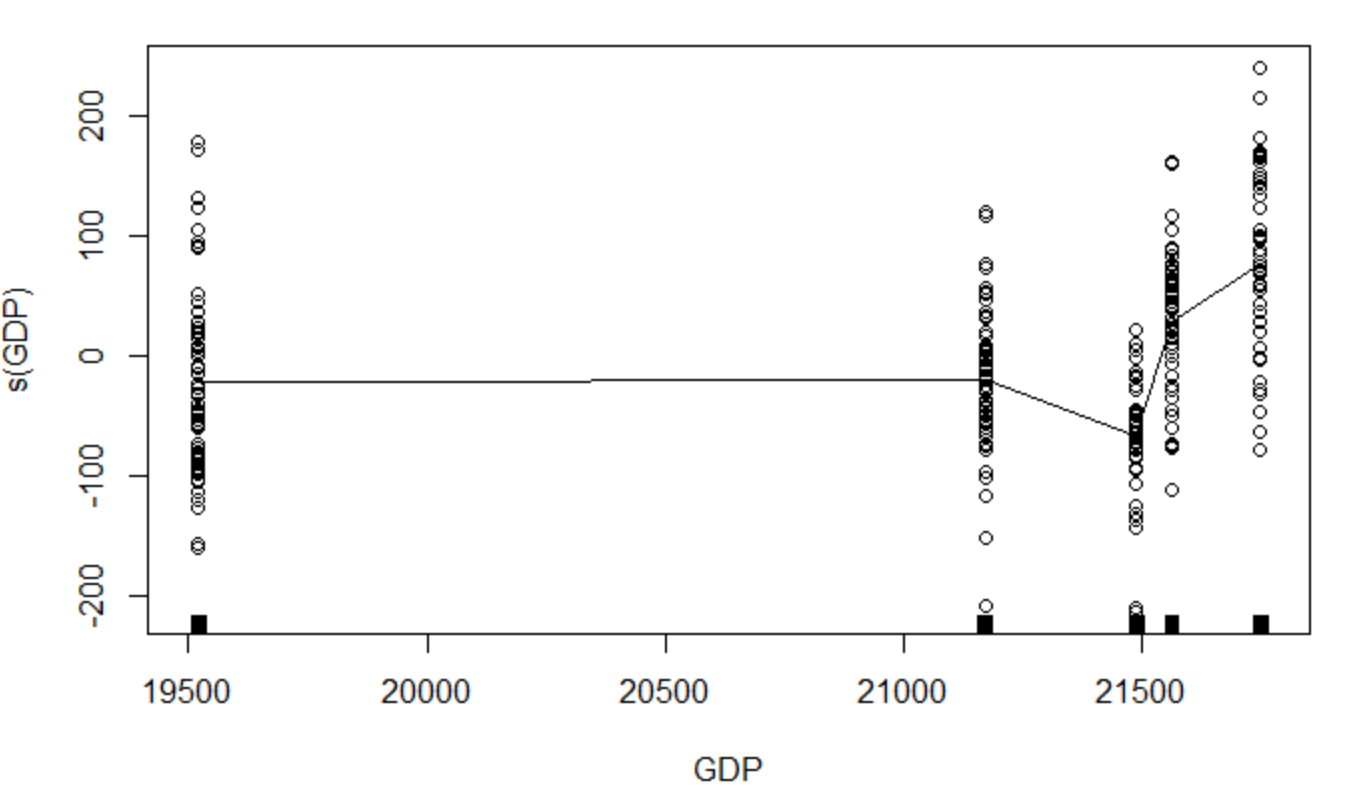
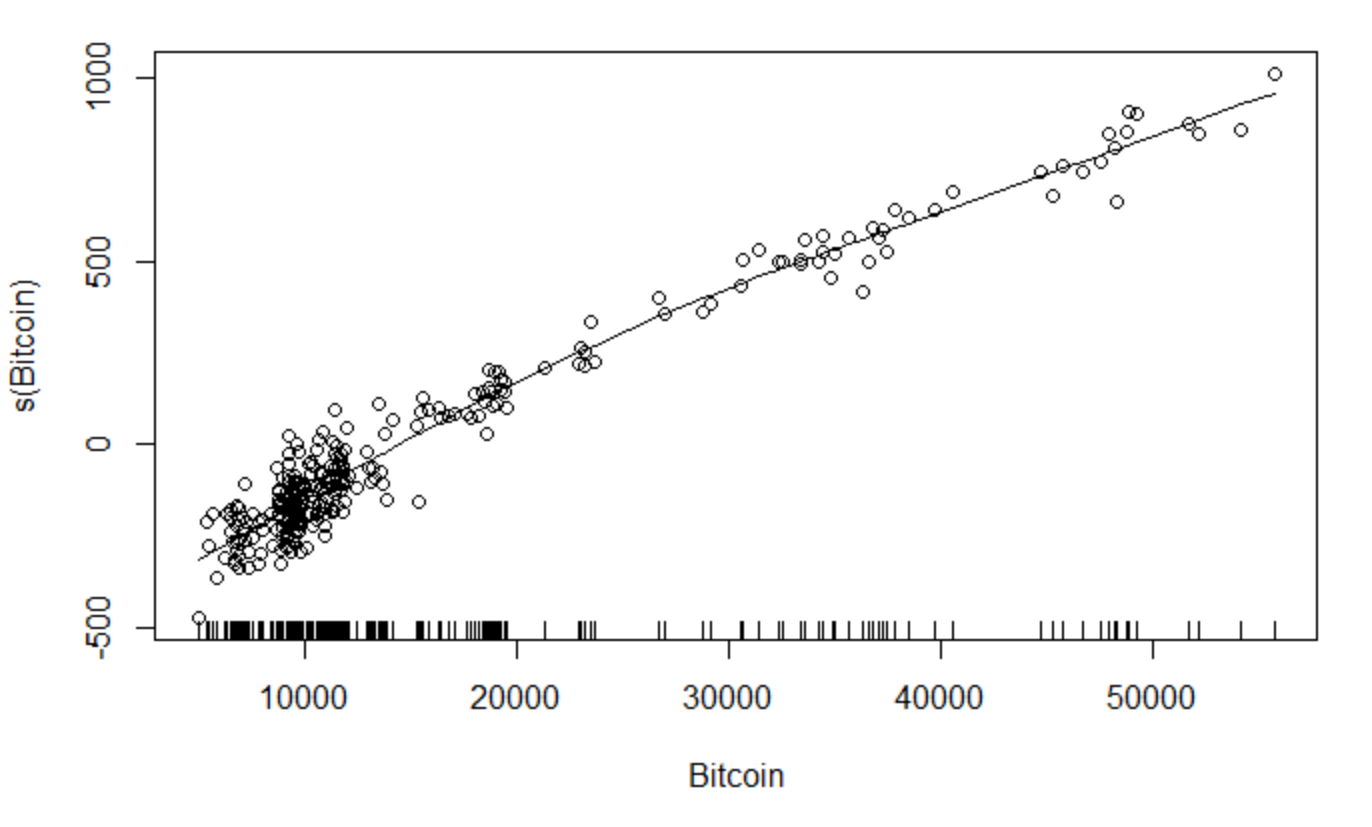
Plots









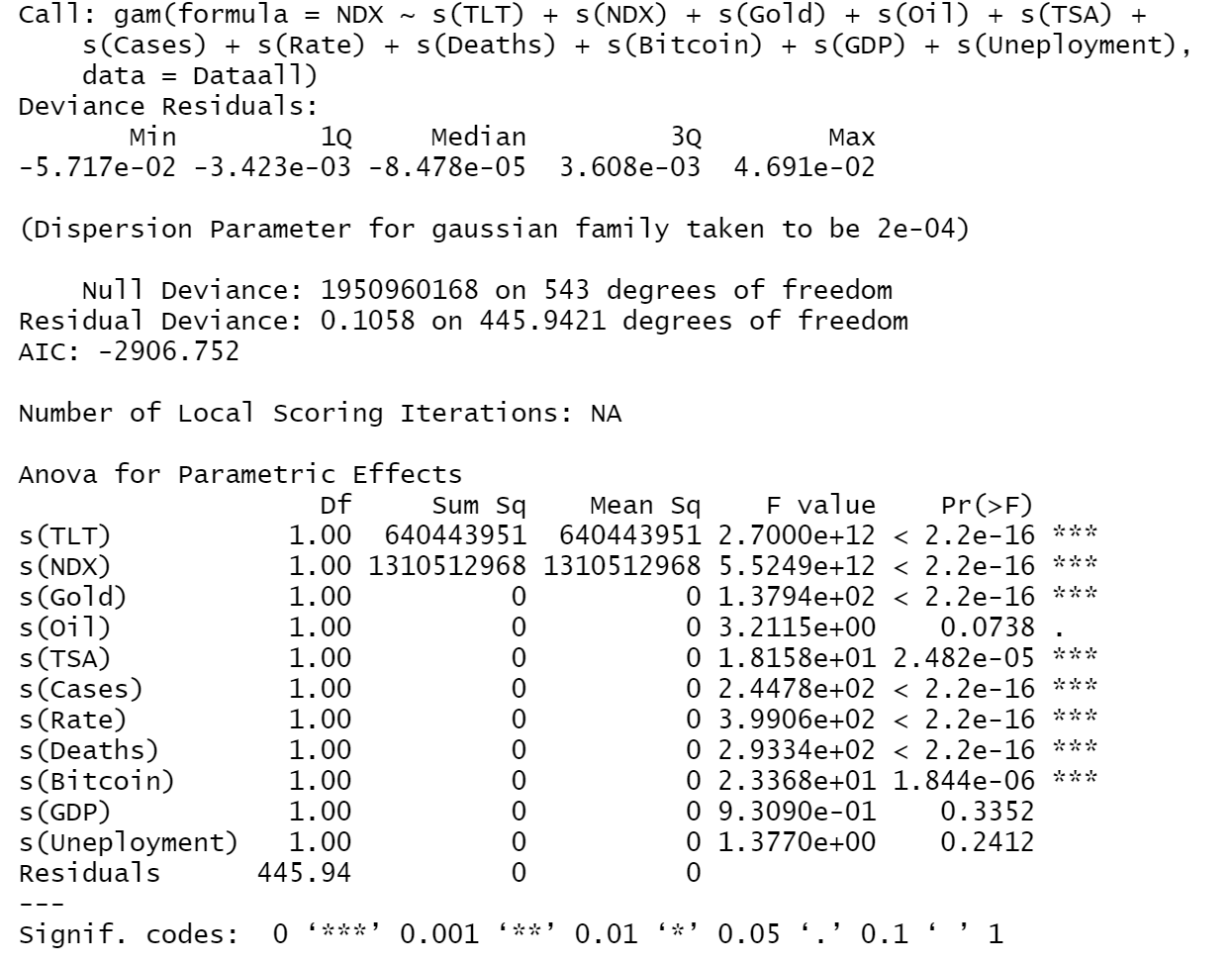




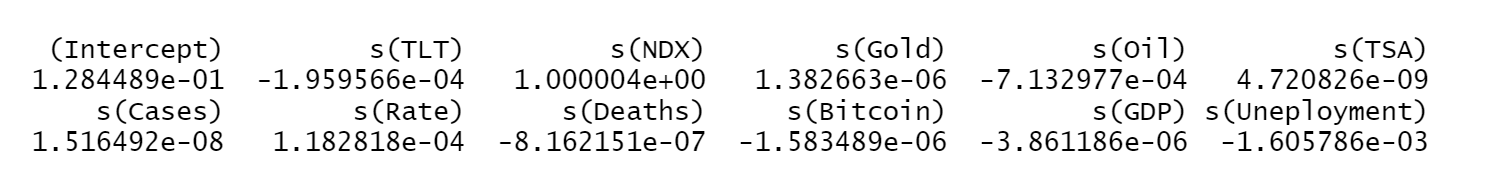
We can see after comparing the two data models for the NDX index that there are notable differences in a couple of variables when comparing the COVID only data and all of the data. Some of the variables that changed were TLT, Oil, TSA, GDP, and Unemployment. These variables are highly correlated to macroeconomic data. During the pandemic bonds were affected due to rate cuts, and oil was affected due to demand across the country. GDP, unemployment, and TSA data were affected due to global lockdowns. Rate, deaths, and bitcoin were not significant in either model. Oil, GDP, and unemployment were significant in the first model but not the Covid only data.

1. **GAM Model DJIA All Data**

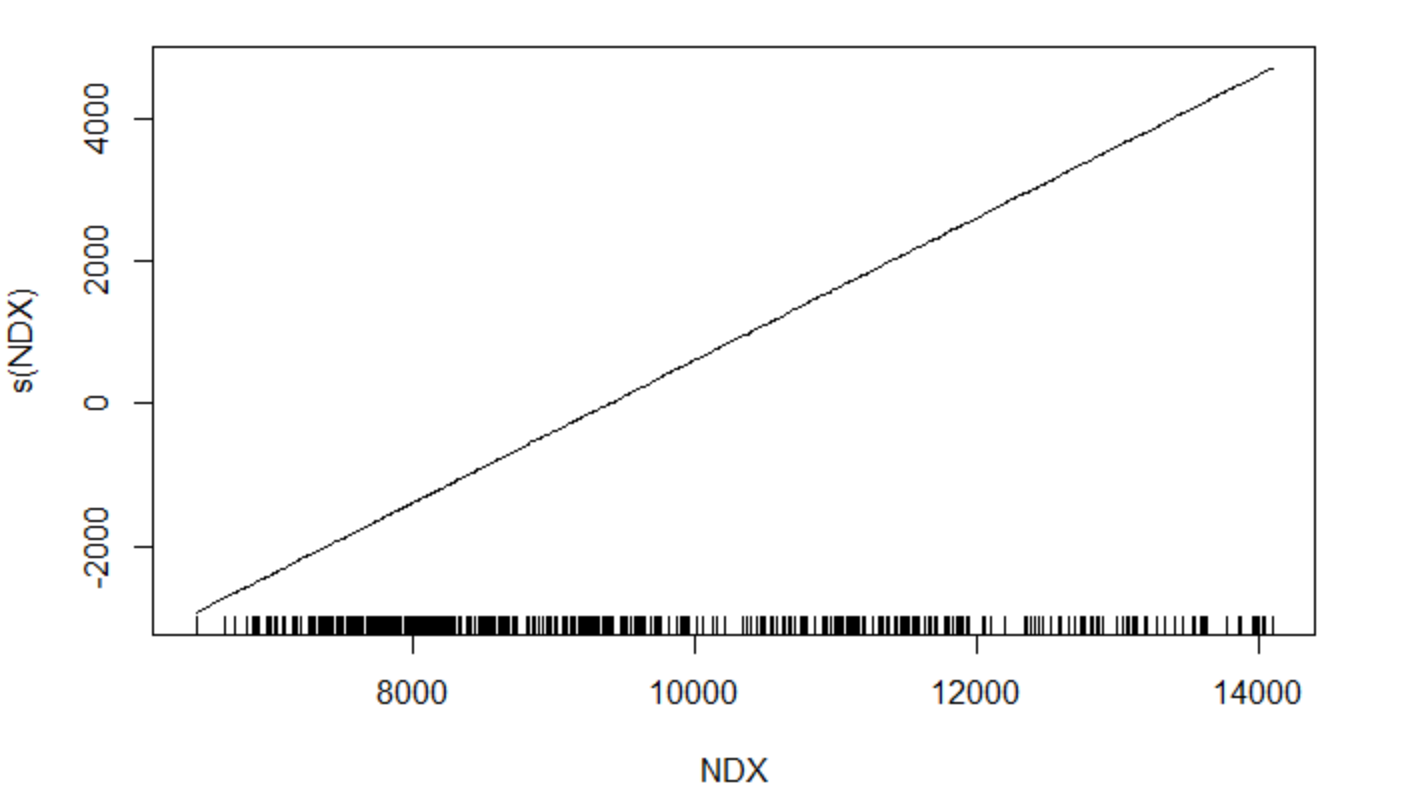
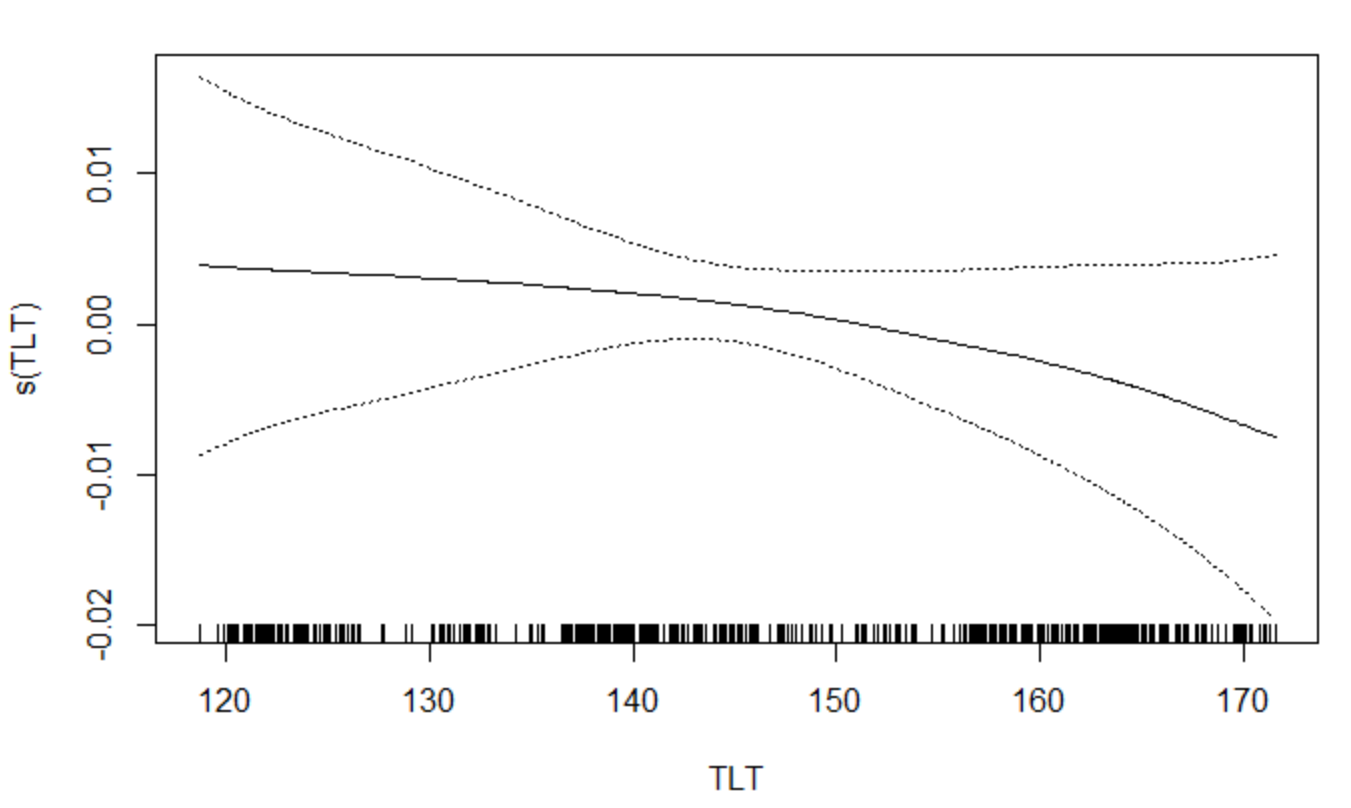
Summary

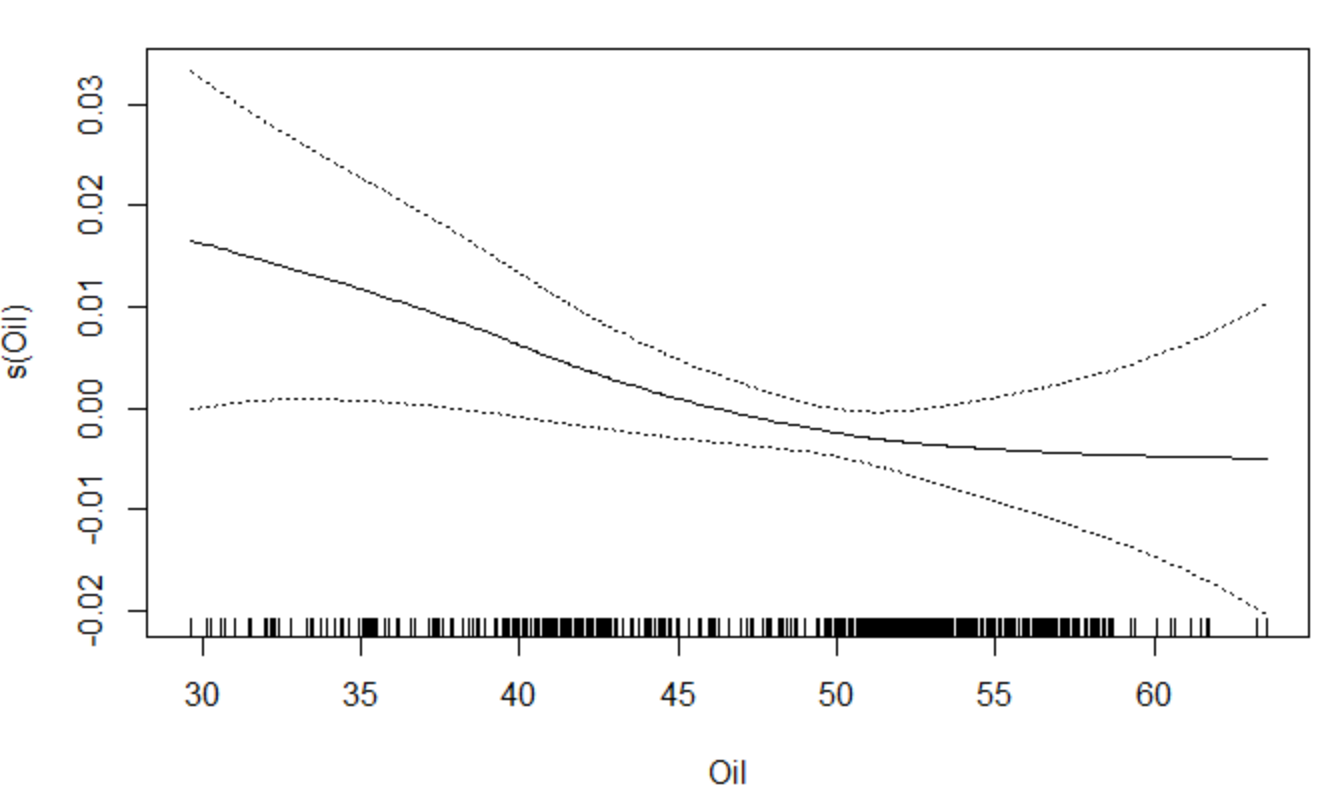
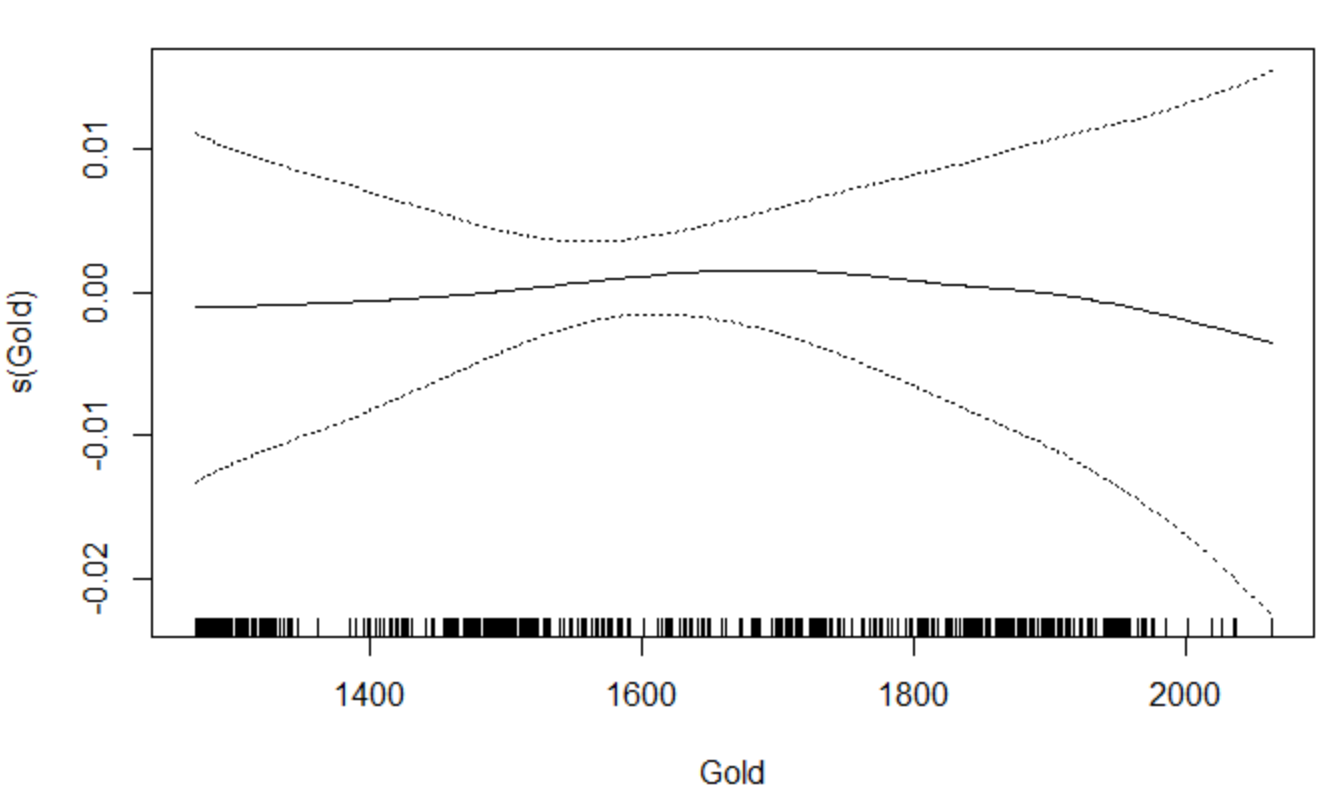
****

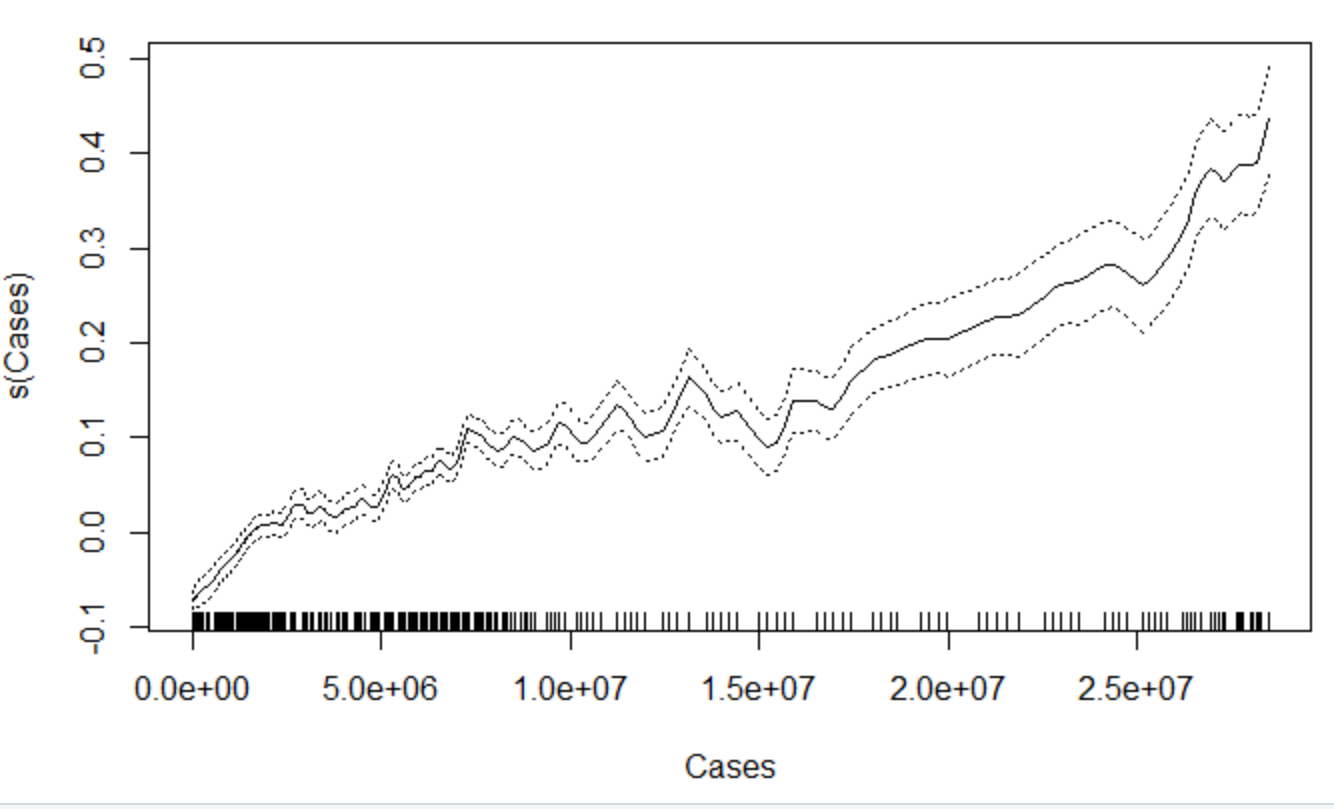
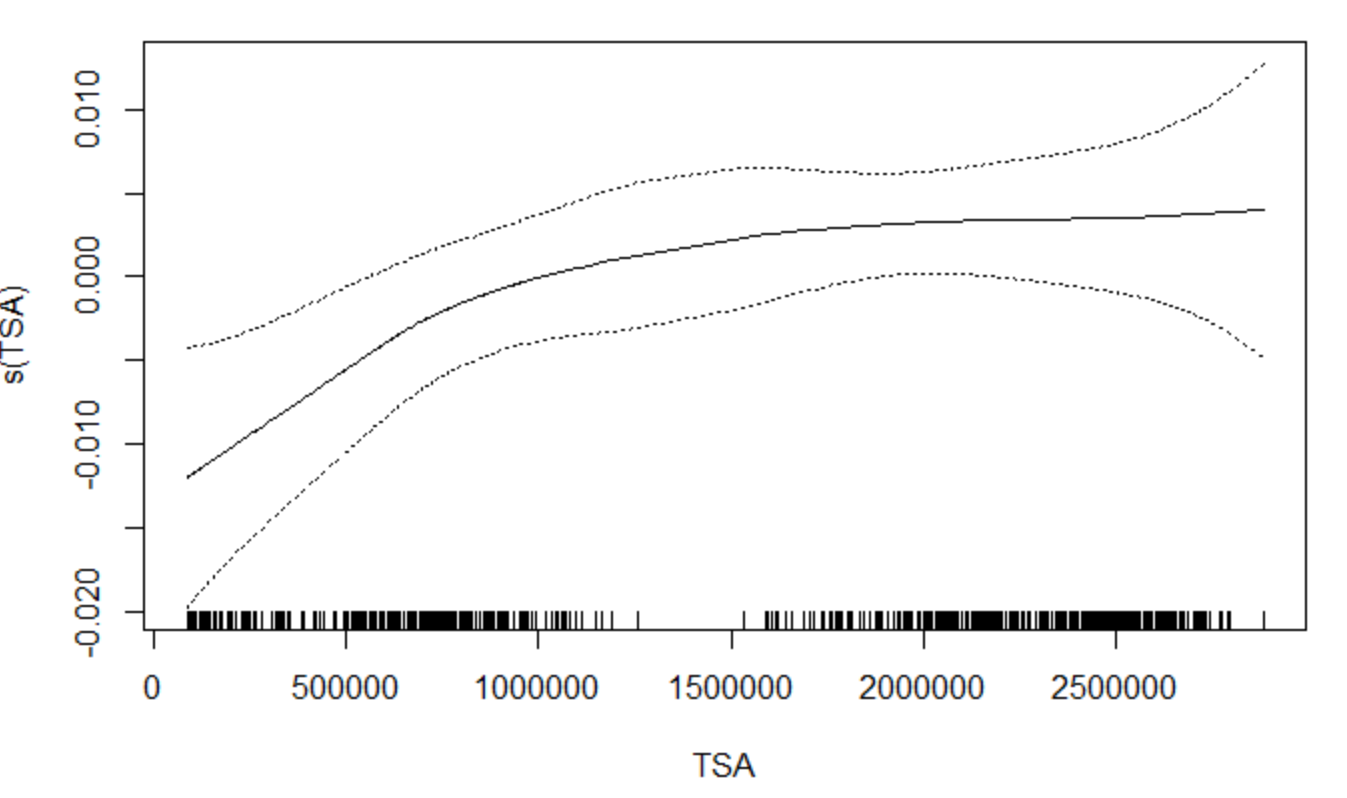
Coefficients

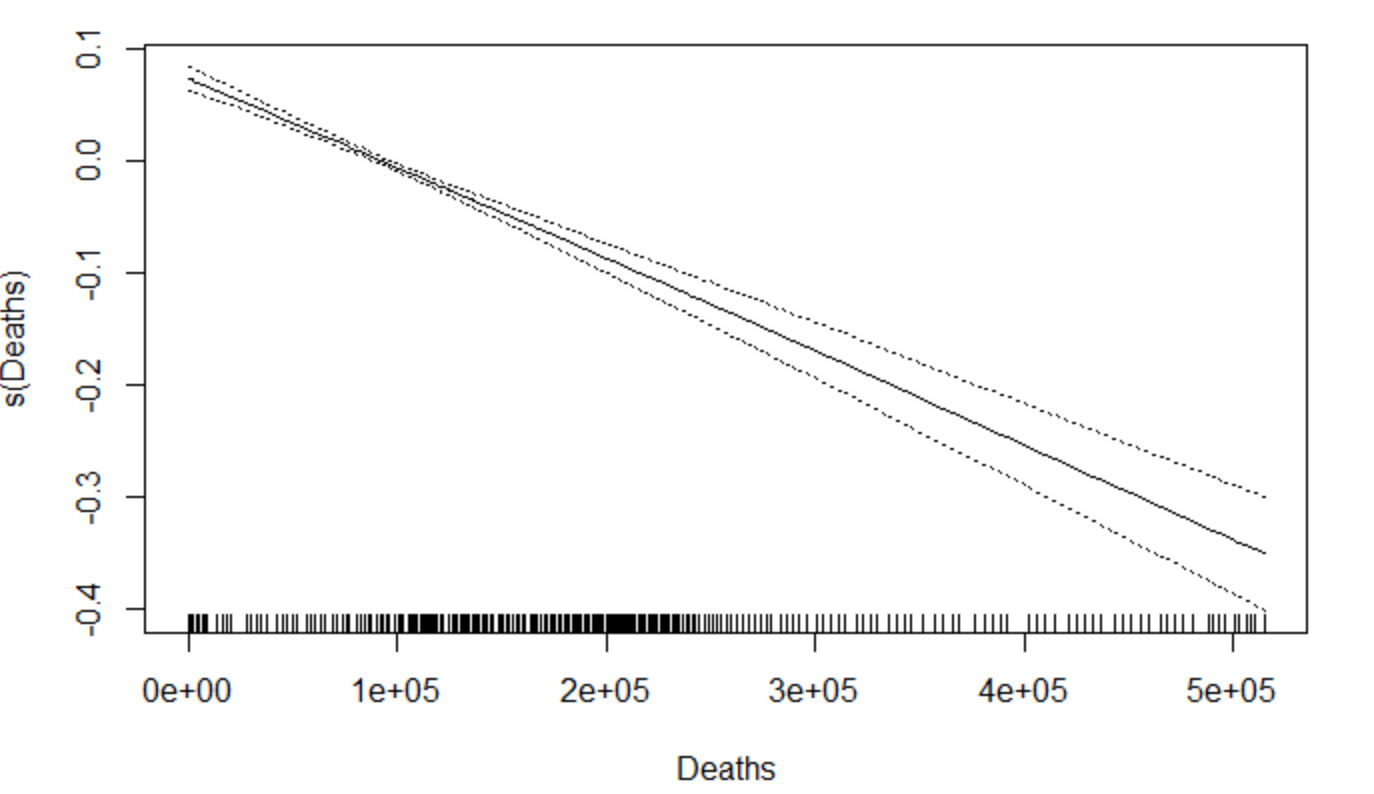
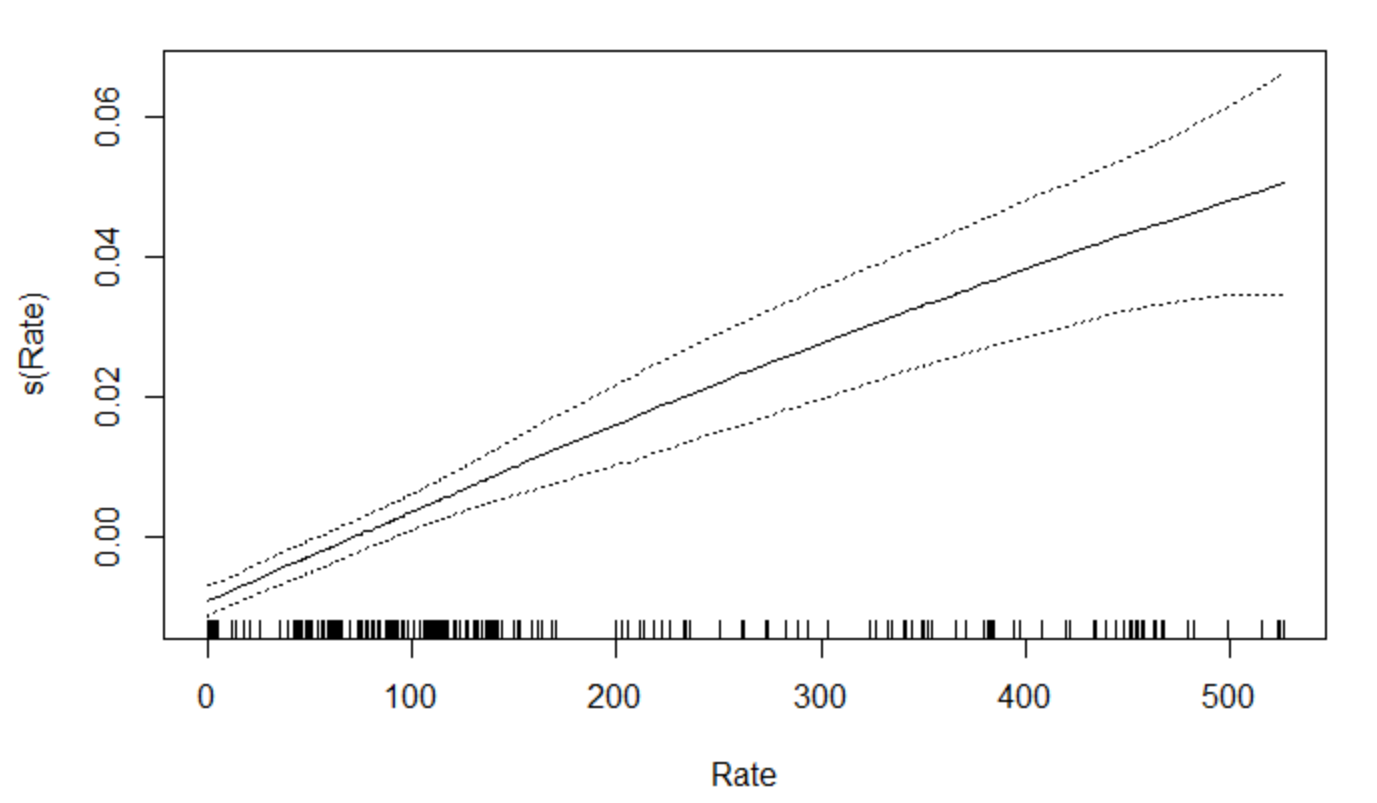


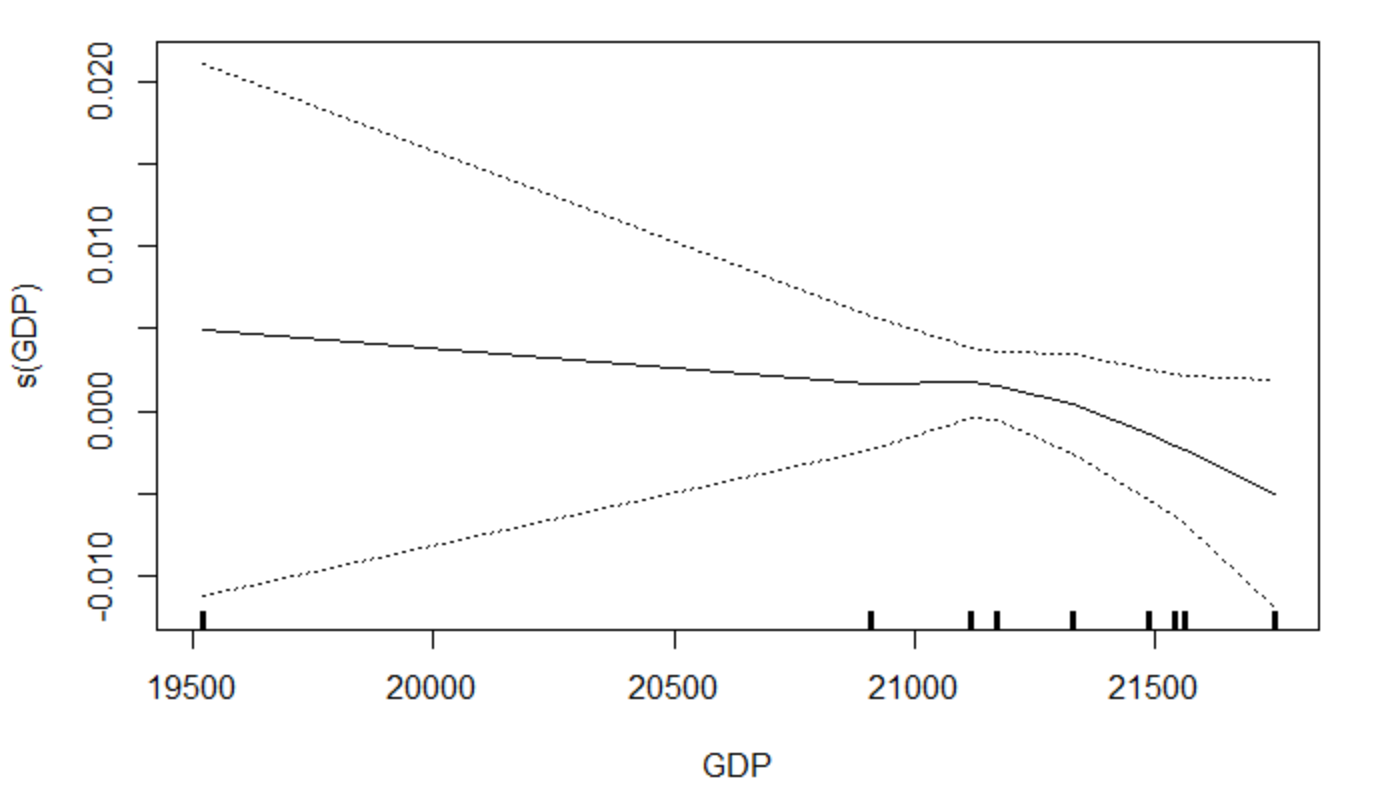
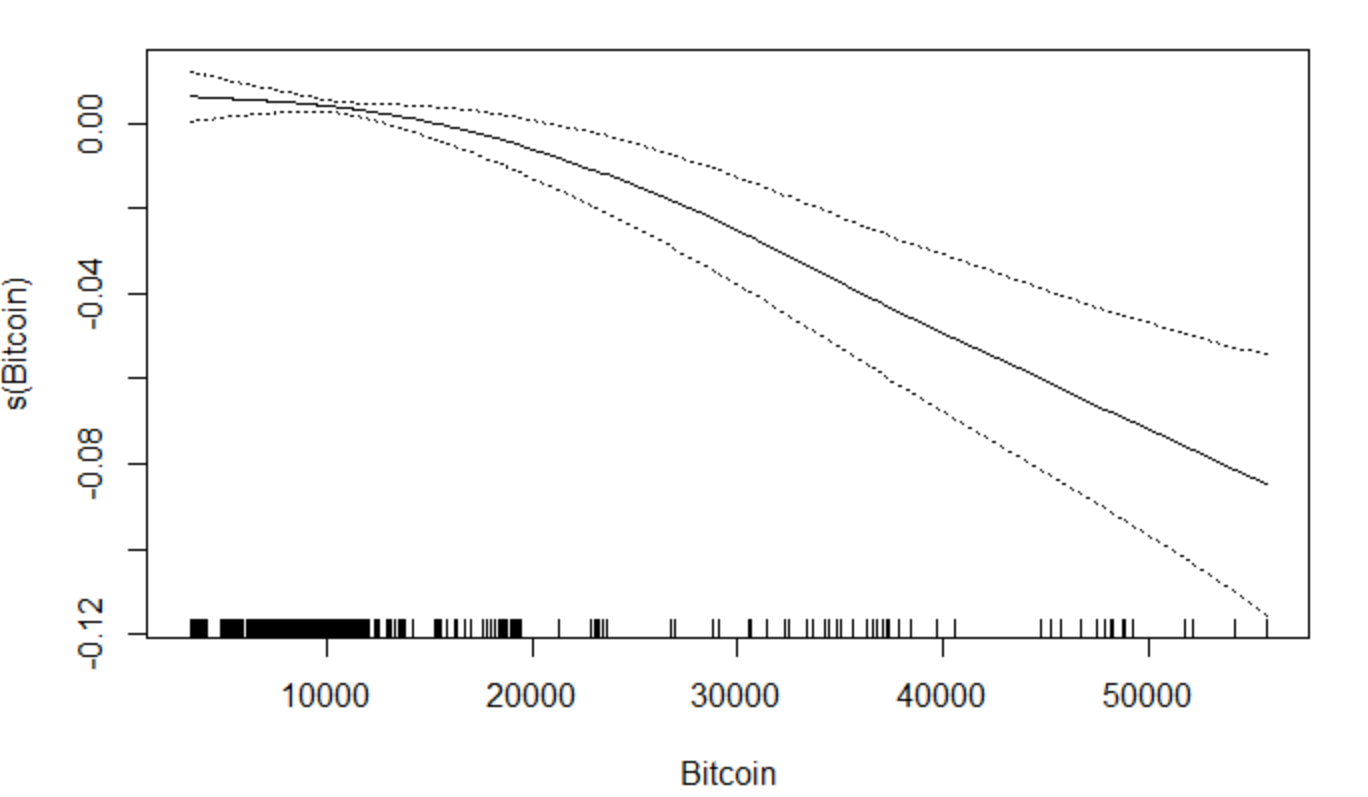
Plots

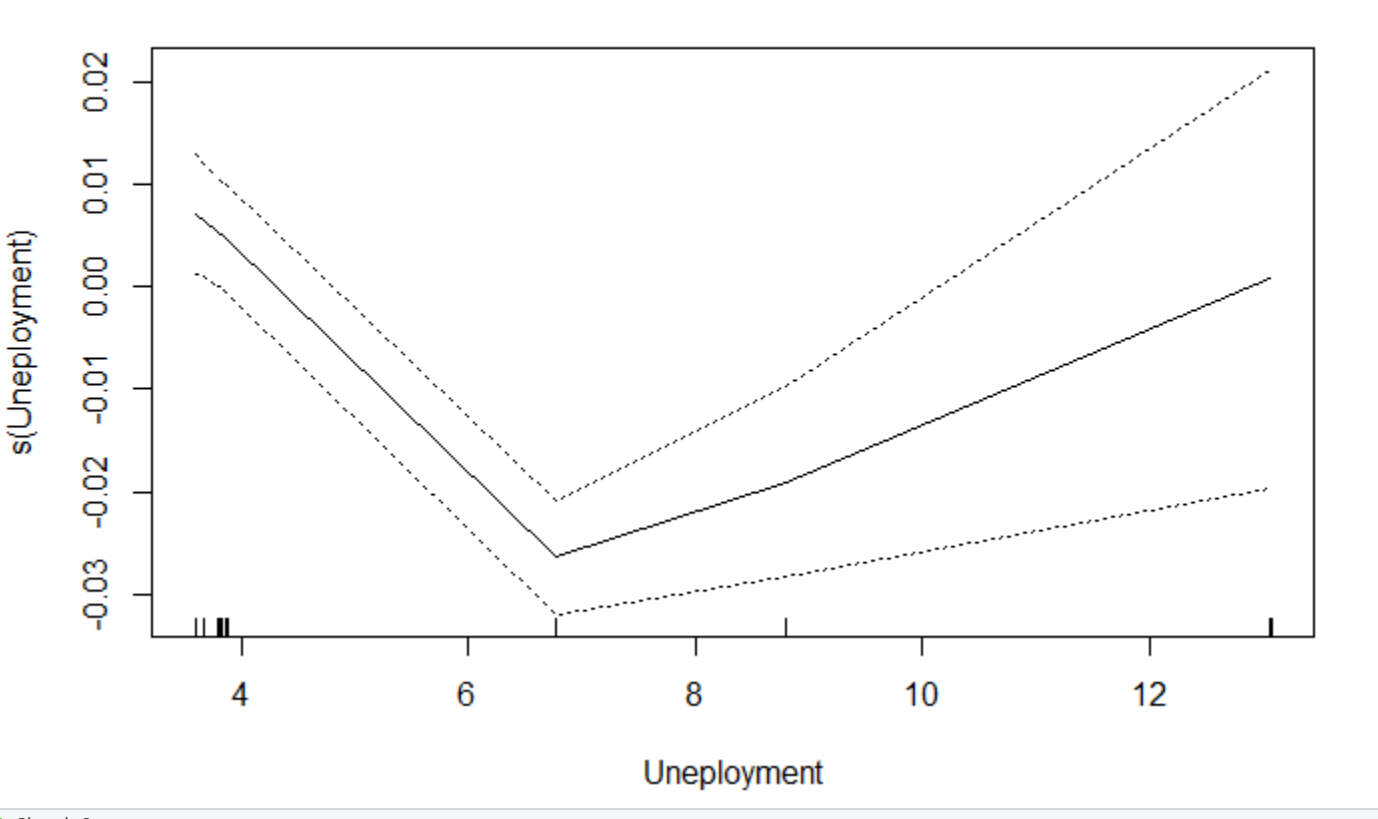






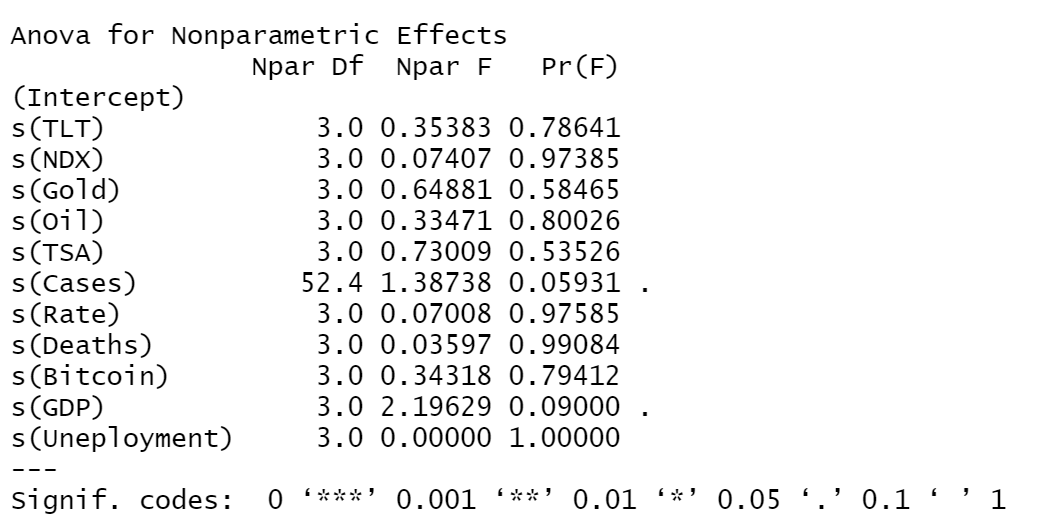
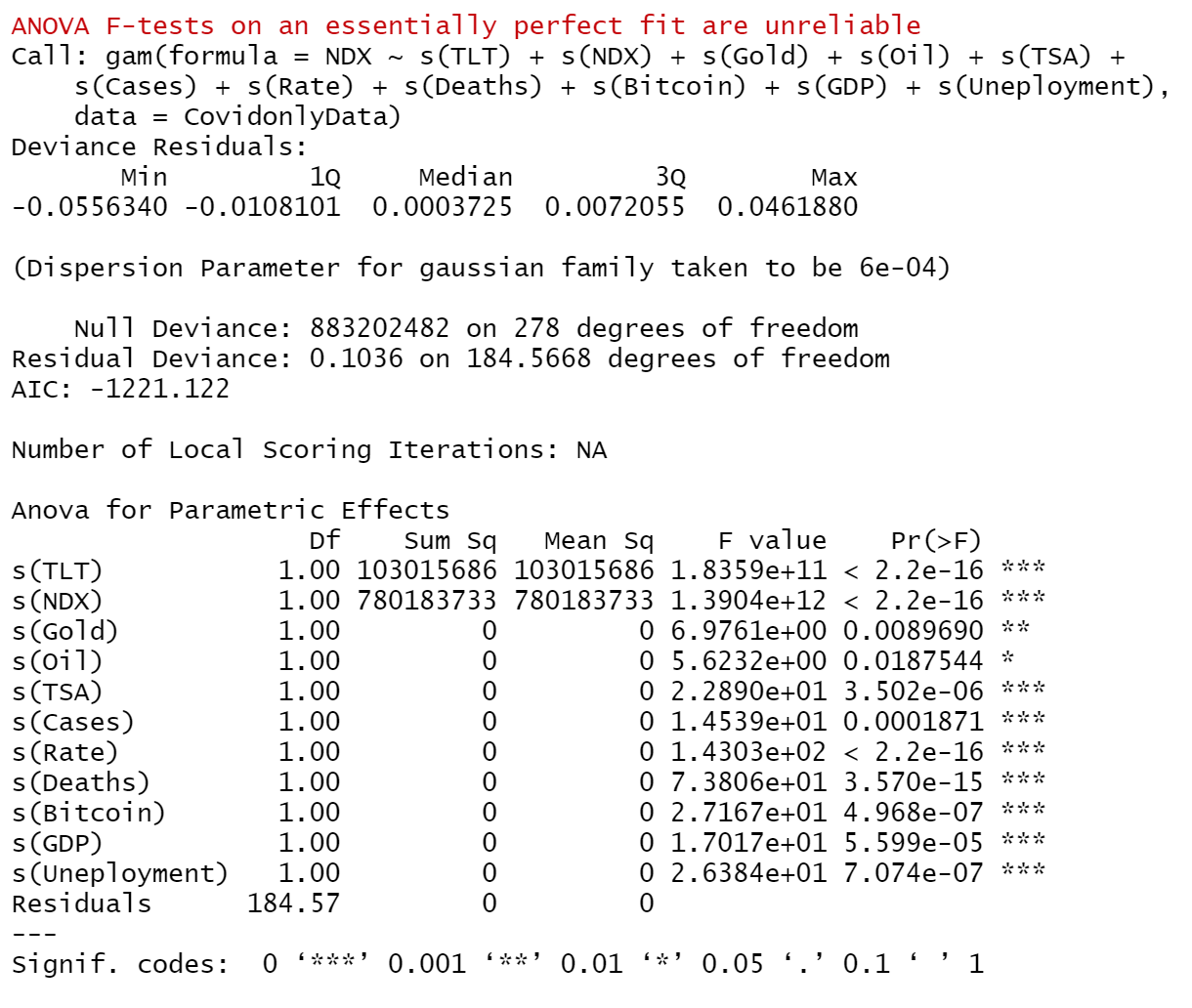




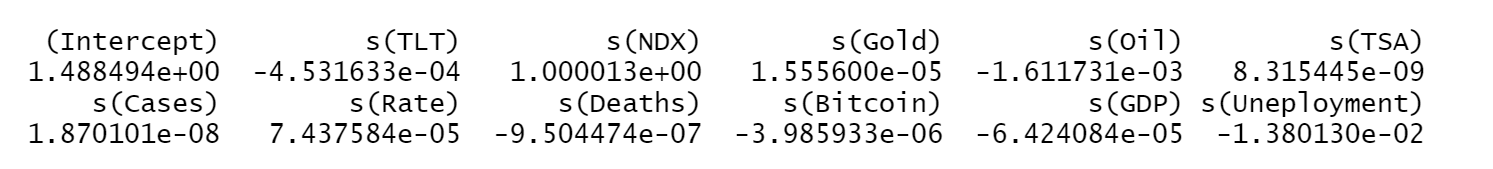


1. **GAM DJIA Covid Only**

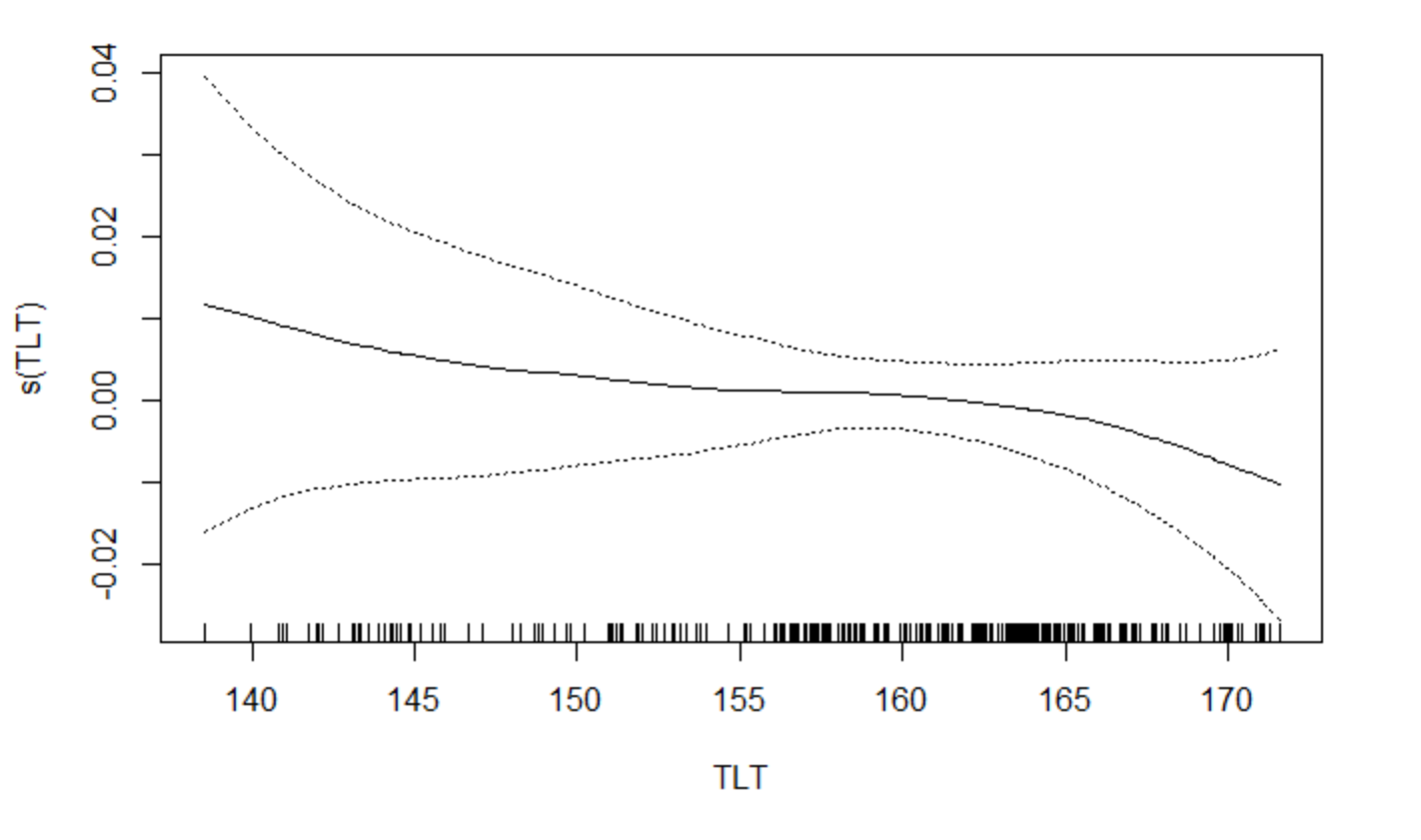
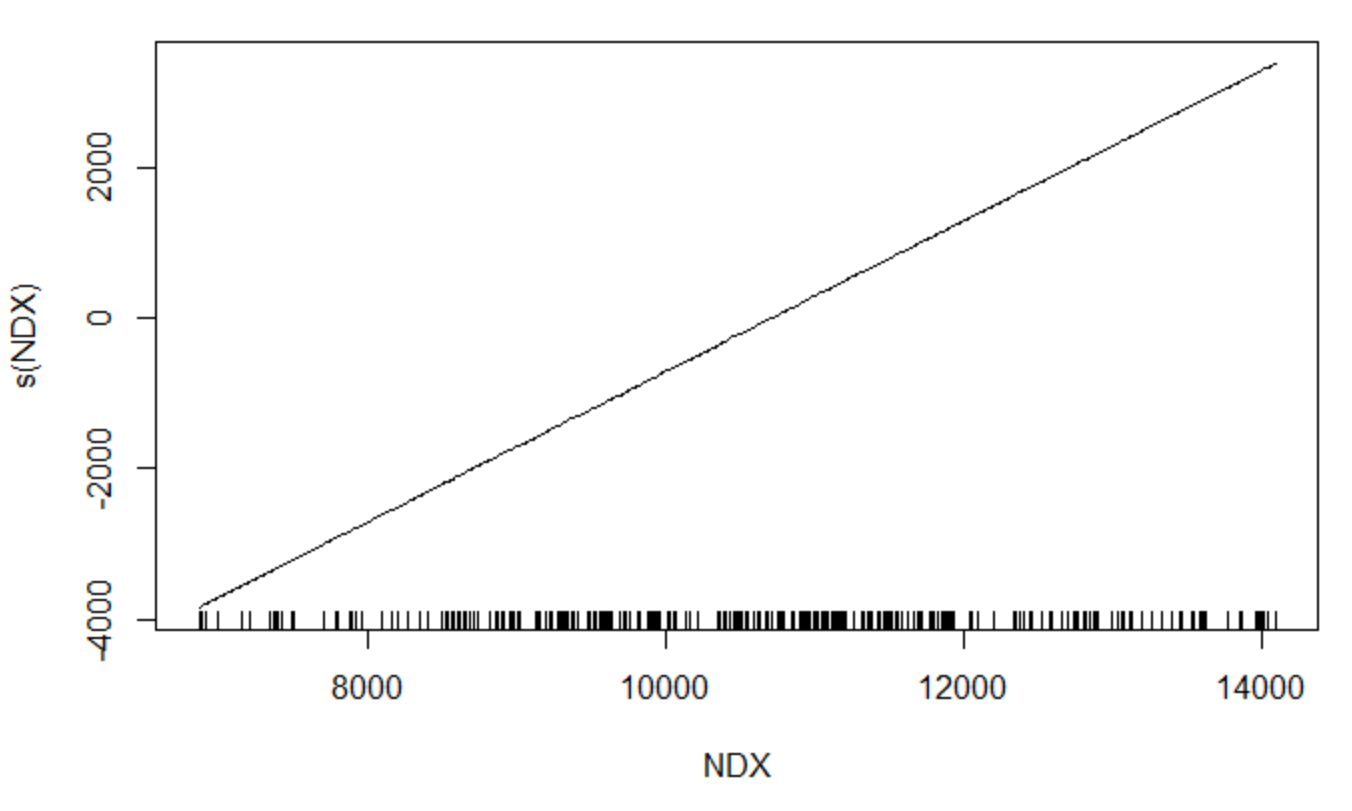
Summary

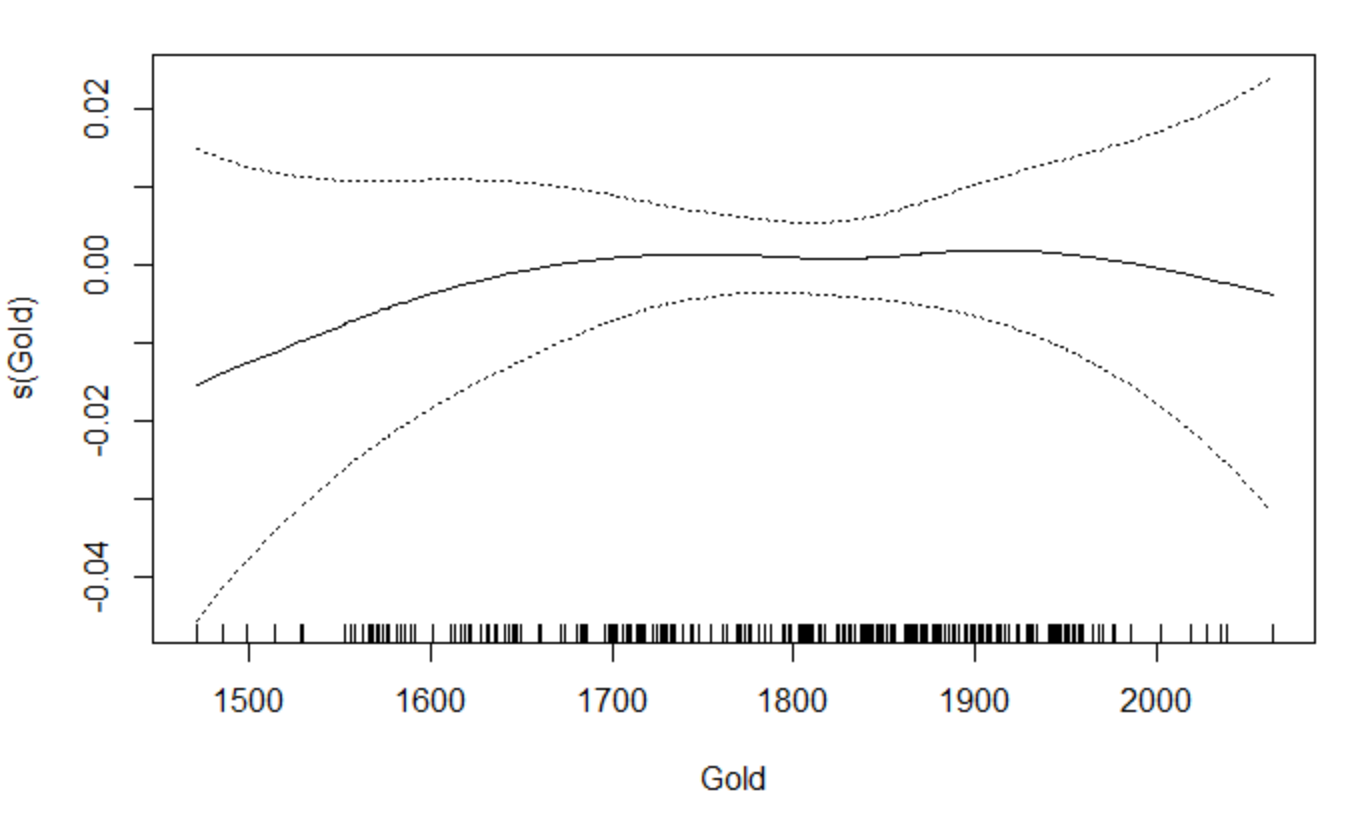
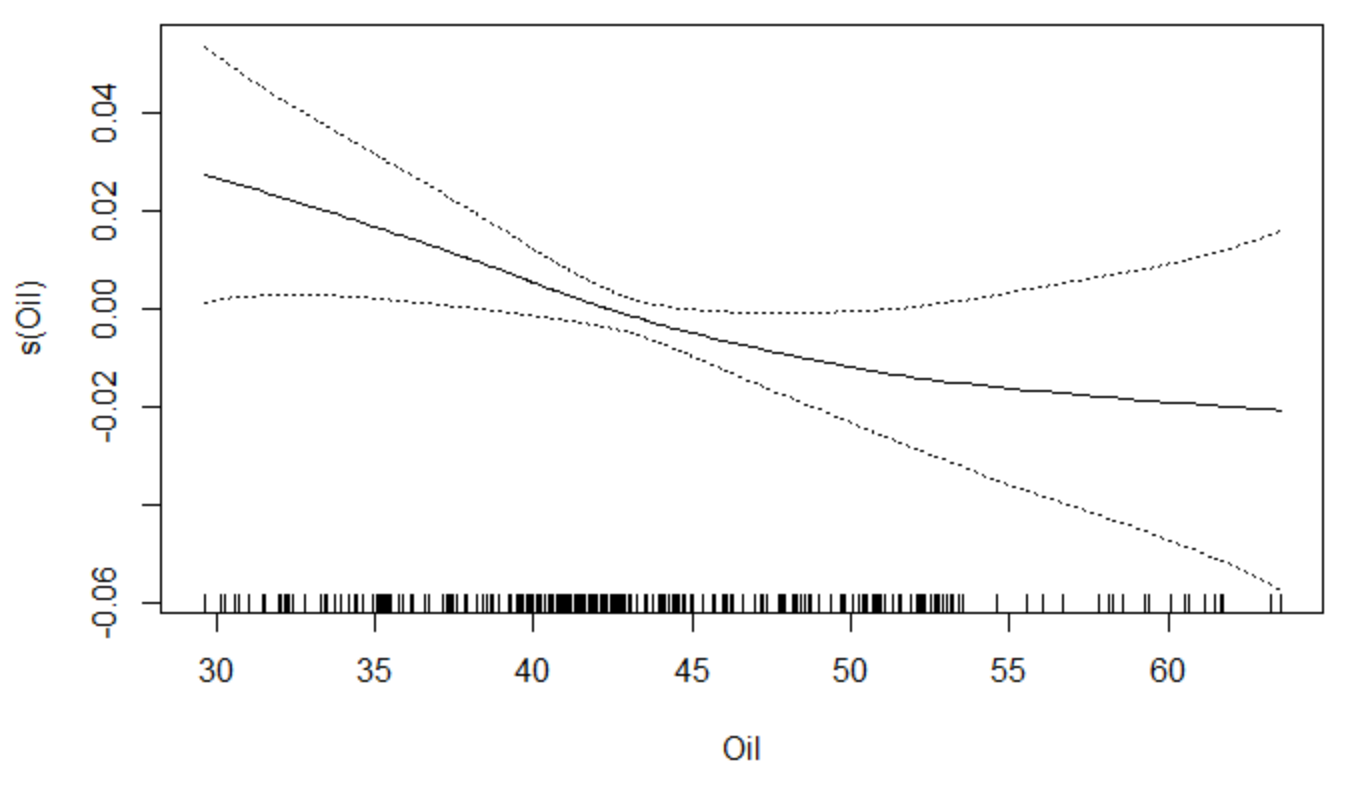


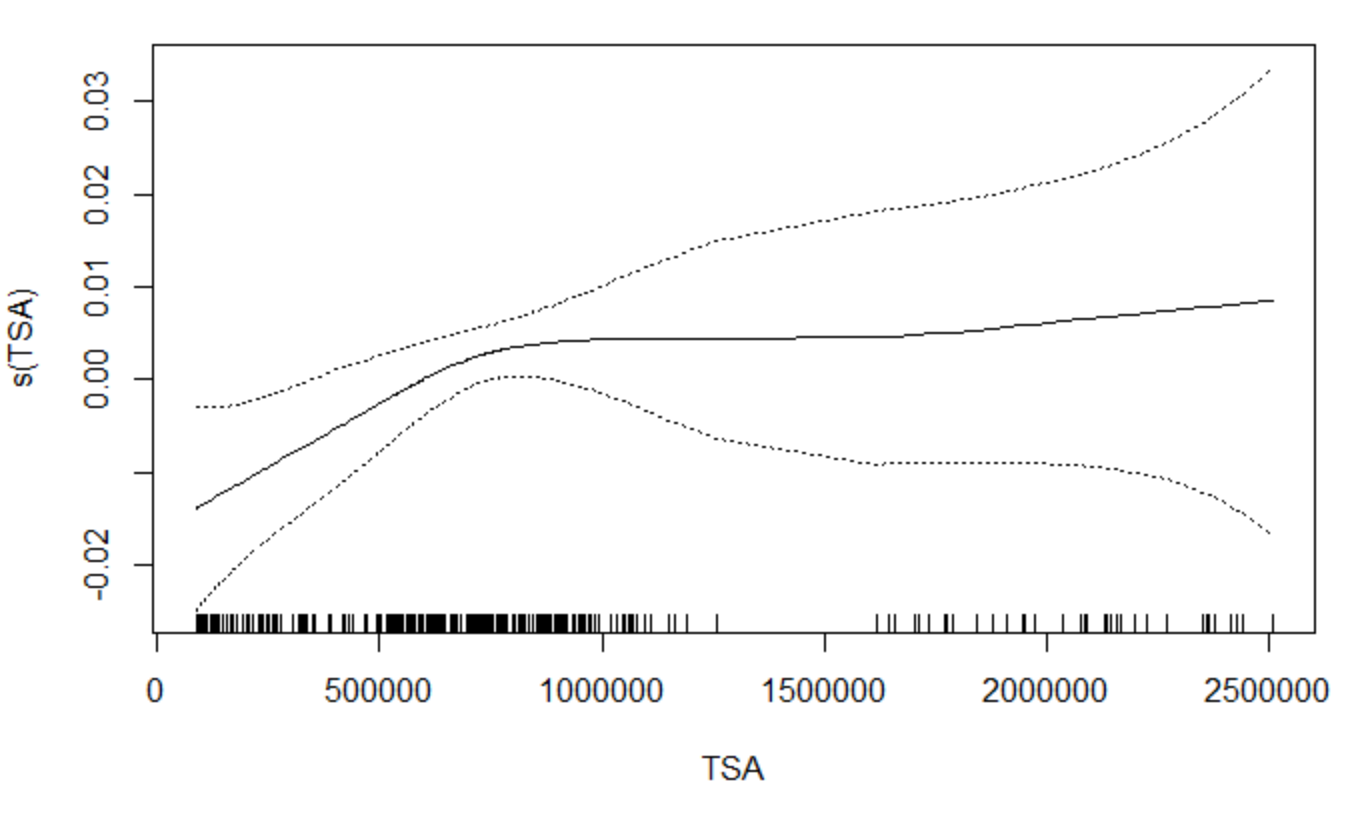
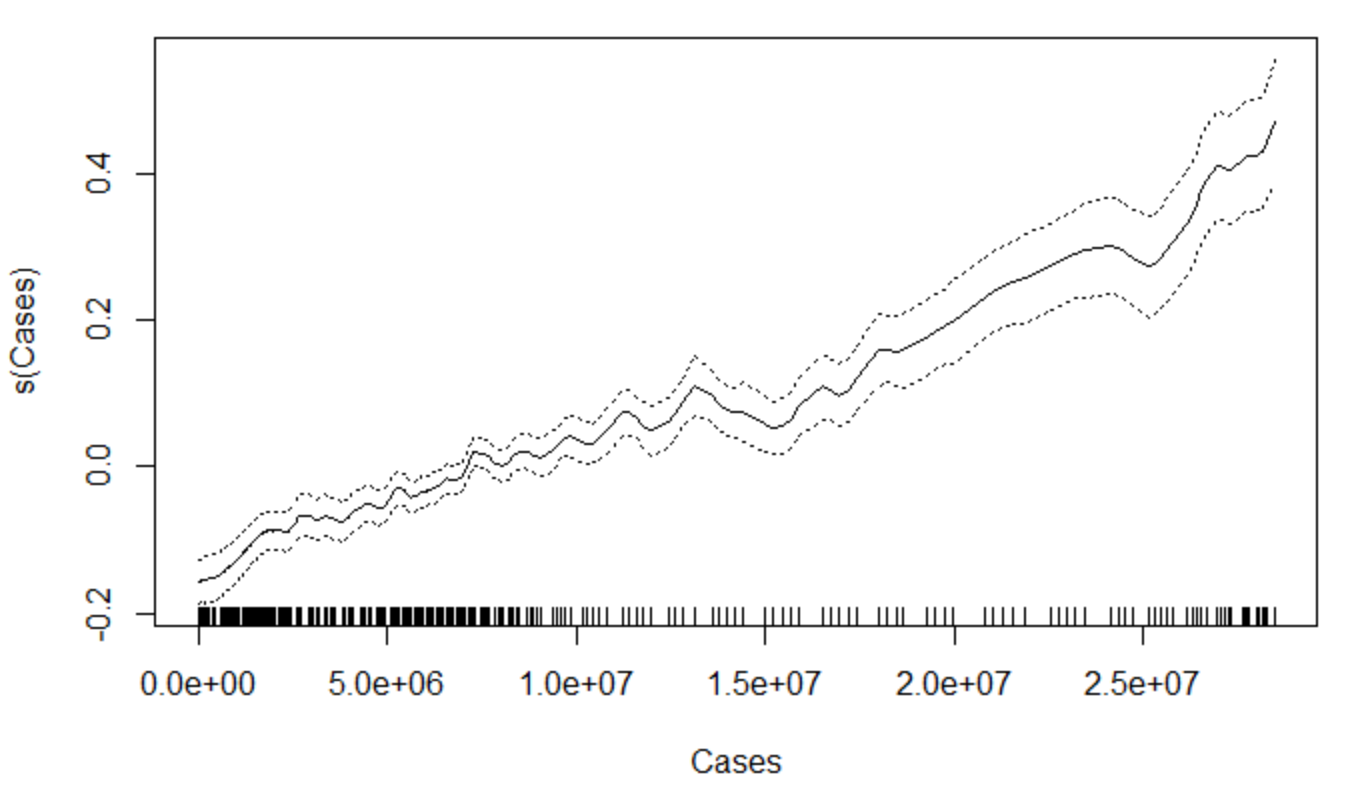
Coefficients

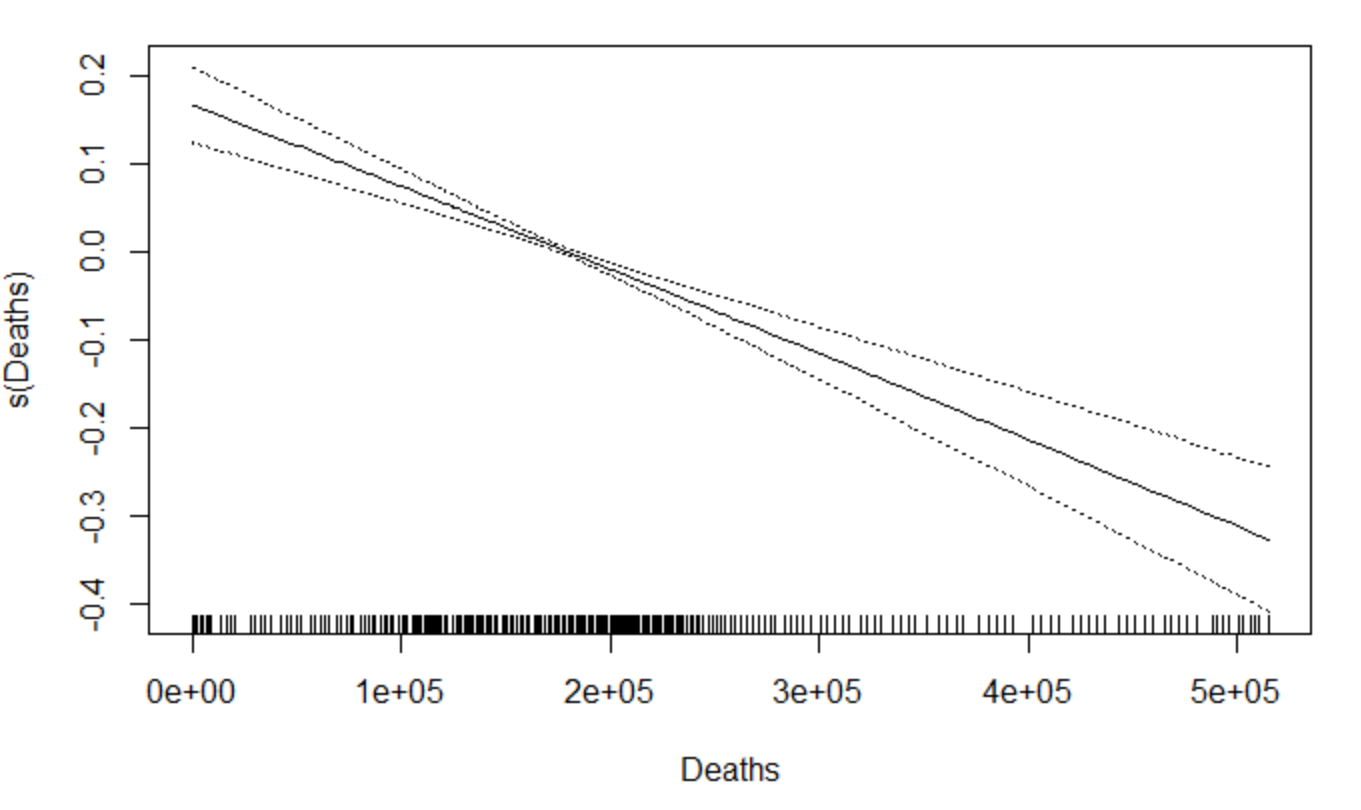
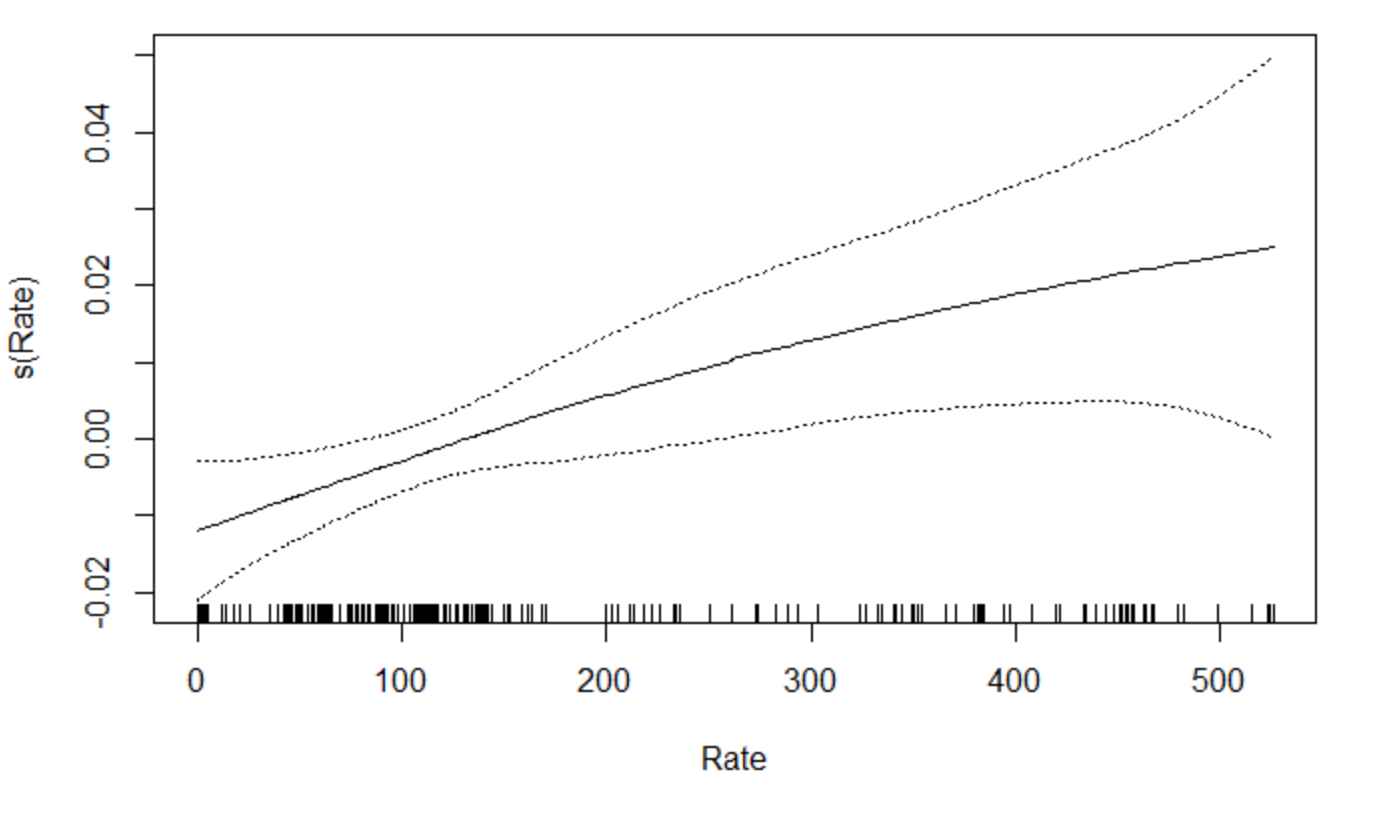


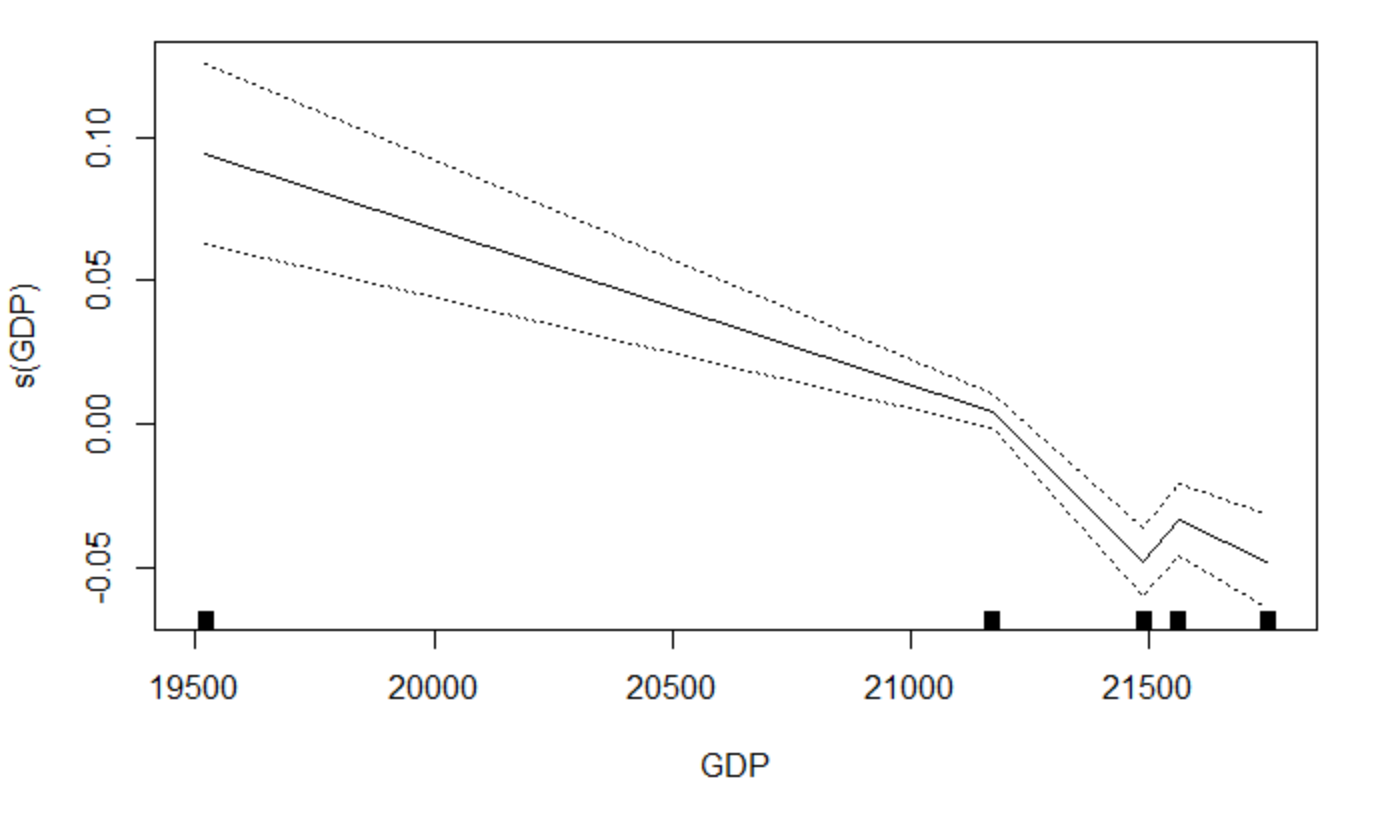
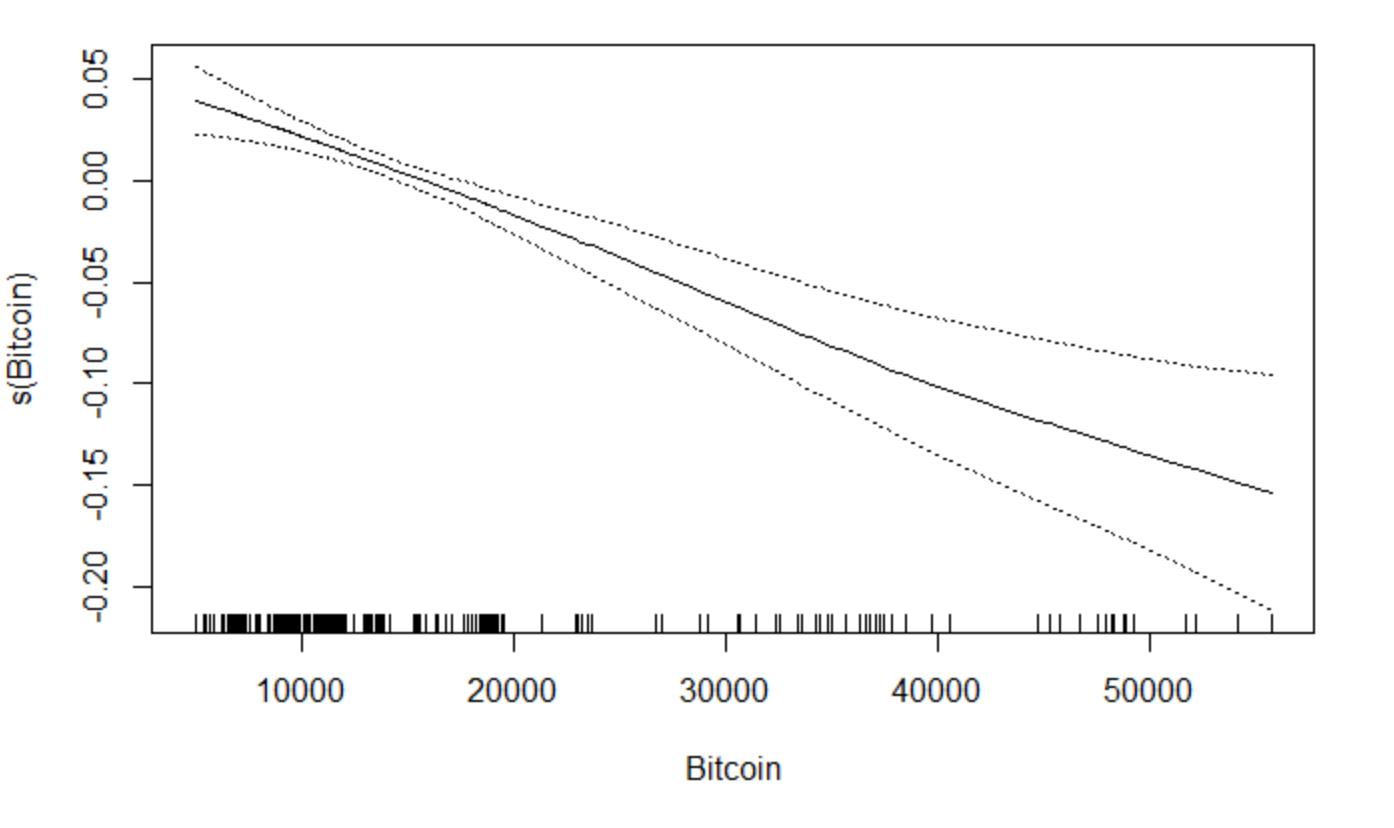
Plots

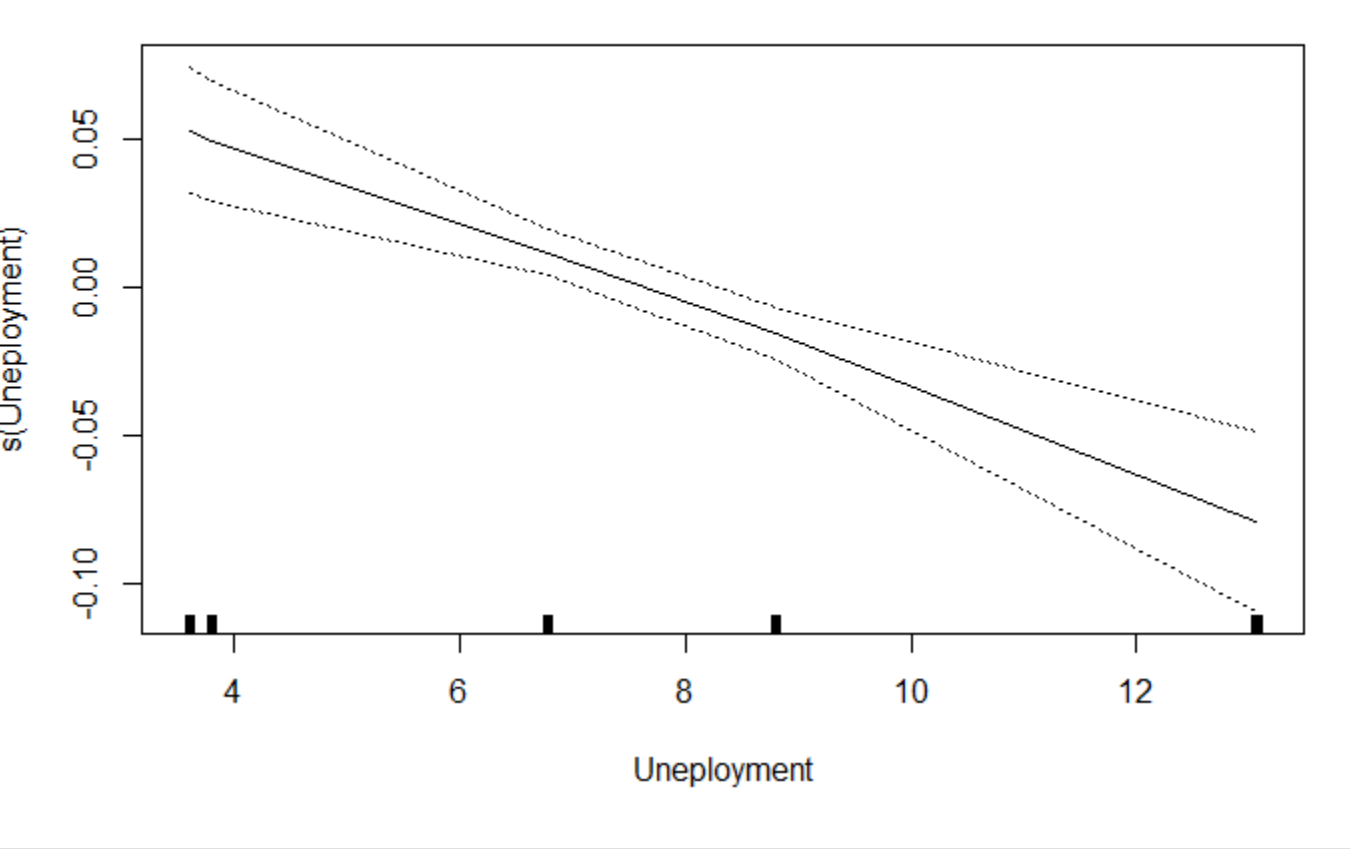
 







When comparing the two DJIA models there were once again a couple of variables that changed shape when comparing the Covid only and all data set. Unemployment appeared to be the one variable that changed the most between the two sets. In the covid only data, there appeared to be no significant variables however in the all data set cases,bitcoin, and unemployment were significant. Over the GAM model might not be the most ideal model in order to analyze and predict future prices.

**Issues with the GAM model**

There appears to be an issue with the modelling of the DJIA model. The variables in the model appeared to have a perfect fit for the anova F test. This means that the variables may be a perfect fit so there are potential issues when it comes to determining which variables are significant in the model.

Random forest

Next model that we evaluated was a nonlinear tree based model, we chose to evaluate random forest with mtry=4 for SPX, BTC, NDX and DJIA. We split the data into test and training sets. Train data date range : 2019-01-02 to 2021-01-10, Test data date range: 2021-01-11 to 2021-03-01

SPX

|  |  |
| --- | --- |
|  | |
|  |  |
|  | |

Model has mean square error of 11653.34 for SPX prediction

Bitcoin

|  |  |
| --- | --- |
|  | |
|  |  |
|  | |

Model has mean square error of 96203861

Dow Jones

|  |  |
| --- | --- |
|  | |
|  |  |
|  | |

Model has mean square error of 259140.1

NASDAQ 100

|  |  |
| --- | --- |
|  | |
|  |  |
|  | |

Model has mean square error of 499300.2

Random forest seems to perform better for SPX than for BTX, DJIA and NDX. change in mtry parameter doesn’t seem to affect the MSE

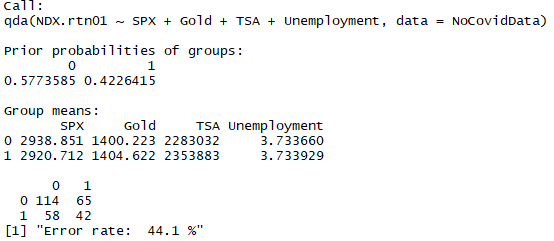
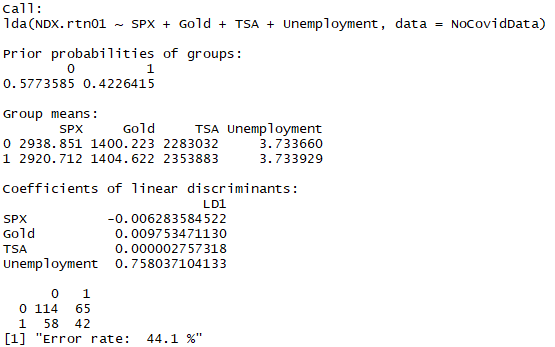
**Linear Discriminant Analysis, Quadratic Discriminant Analysis and K-Nearest Neighbors**

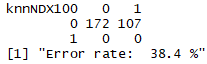
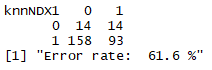
LDA, QDA and KNN were evaluated as models for this data. It was hypothesized amongst the team that these models may not give the best results, however they were worth ruling out. As a test, these models were created to predict whether log return of NDX would go up or down. These models were created three ways:

1. Using only pre-Covid data as a training set
2. Including Covid data in the training set
3. Including Covid data in the training set and only using Covid variables as predictors (i.e. Cases, Rate and Deaths)

After exploring the data and evaluating the correlation matrix, NDX was modeled based on SPX, Gold, TSA and Unemployment variables; with the addition of Deaths and Cases when Covid data was included in the training set.

**LDA, QDA and KNN trained on pre-Covid Data**

****

****

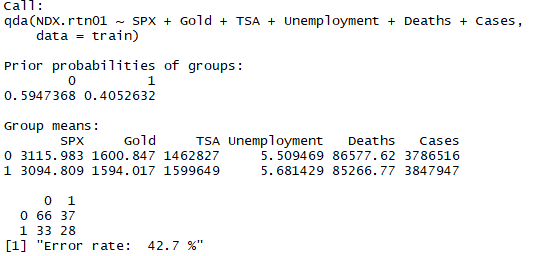
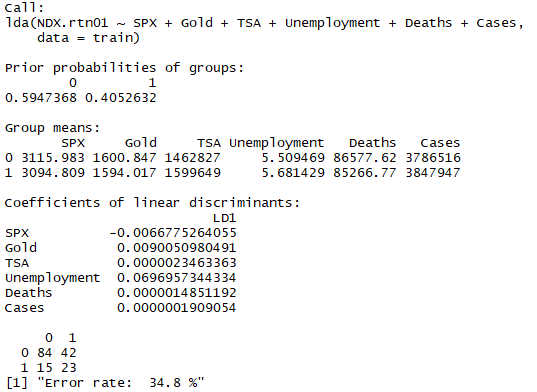
LDA error rate = 44.1%

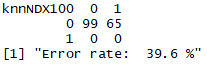
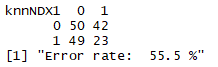
QDA error rate = 44.1%

KNN (k=1) error rate = 61.6%

KNN (k=100) error rate = 38.4%

**LDA, QDA and KNN trained including Covid Data**

****

****

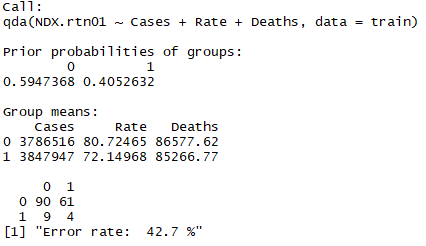
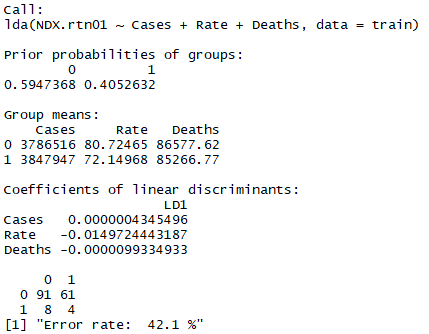
LDA error rate = 34.8%

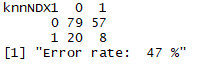
QDA error rate = 42.7%

KNN (k=1) error rate = 55.5%

KNN (k=100) error rate = 39.6%

**LDA, QDA and KNN trained on Covid Data Variables Only**





LDA error rate = 42.1%

QDA error rate = 42.7%

KNN (k=1) error rate = 47%

KNN (k=100) error rate = 39.6%

**LDA, QDA and KNN Conclusions**

The error rates for all of the models were not good. Most of the models had error rates that were in the 40-50% range with a few outliers The lowest was LDA including Covid data in the training set at 34.8%. The highest error rate was KNN (k=1) trained on pre-Covid data, at 61.6%, that mostly predicted NDX to increase in return. Like models seem to reduce their error rate when incorporating Covid data. There does not seem to be an improvement with only using Covid variables, however. One obvious factor that is not evaluated as a predictor in these models is simply the Date. As time increases NDX increases overall. One could simply predict that as the Date increases, NDX will increase. This may be true based on history and in the order of magnitude of decades, given the history of the stock market in general; however forecasting based only on the Date would lead to very significant errors in smaller time windows (there have been extended and significant downturns in the market before). Another possible conclusion is that Covid did not have an impact on the overall market nor NDX as a baseline for the market. Either way, LDA, QDA and KNN do not seem to be good models for this data.

**Methods:**

Classification

1. GAM
2. Random Forest
3. LDA/QDA/KNN

**Next Steps:**

The next steps for this project is to determine which of the following models to use going forward and which variables need to be included/excluded in the model. After this the next step would be to project future prices for a portfolio based on past data and how future variables may change in the future.

**References:**

[1]The relationship between cryptocurrencies and COVID-19 pandemic<https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7388435/>

[2] Impact of COVID-19 on forecasting stock prices: An Integration of Stationary Wavelet Transform and Bidirectional Long Short-Term Memory <https://www.hindawi.com/journals/complexity/2020/1846926/>

[3] Bloomberg Terminal

|  |  |
| --- | --- |
| **Data** | **Function** |
| TSA Throughput | TSATTPCY |
| CoVID-19 Cases | NCOVUSCA |
| CoVID-19 Recovered | NCOVUSRE |
| CoVID-19 Deaths | NCOVUSDE |
| CoVID-19 Vaccine administered | NCOVUSVA |
| S&P 500 | SPX Index |
| Dow Jones | INDU Index |
| Nasdaq Composite | CCMP Index |
| Gold | XAU Curncy |
| Crude oil futures | C01 Comdty |
| Bitcoin/USD | XBTUSD |
| GDP, Non-farm payroll | ECST |